Quantifying uncertainty in climate-driven disease risk predictions

Thesis submitted on 20th March 2013 to the University of Liverpool
for the degree of Doctor of Philosophy

Dave MacLeod

Supervised by Prof. Andy Morse and Prof. Matthew Baylis

"I have approximate answers and possible beliefs and different degrees of certainty about different things, but I'm not absolutely sure of anything, and many things I don't know anything about...I don't feel frightened by not knowing things, by being lost in the mysterious universe without having any purpose, which is the way it really is, as far as I can tell, possibly. It doesn't frighten me."

Richard Feynman

Abstract

This thesis considers the uncertainty in forecasts of climate-driven disease risk, focusing on seasonal and decadal timescales.

An analysis of the skill of decadal climate predictions is carried out, looking at the first multi-model decadal hindcast set produced as part of the ENSEMBLES project. Some skill in the prediction of global average temperature trends over the forthcoming decade is shown, with no skill evident for precipitation. Focusing on smaller areas shows limited skill in predicting temperature trends and no skill for precipitation trends, suggesting that decadal climate models cannot currently make useful predictions of disease risk.

Seasonal climate forecasting skill is then considered. Seasonal hindcasts produced by two research projects, DEMETER and ENSEMBLES, are compared with the most recent version of the European Centre for Medium-Range Weather Forecast's seasonal forecast model, System 4. The models are validated over Africa and the Indian subcontinent, and it is shown that in general System 4 forecasts are an improvement over the DEMETER and ENSEMBLES multimodel ensembles, particularly for West Africa. A more in depth study of System 4 is subsequently carried out, comparing the variation in skill between forecast start dates. Forecast value is demonstrated at multiple lead times, with most skill found for West African regions and Botswana and limited skill for India; indicating when and where forecasts can potentially be issued to users.

Forecasting malaria is then studied by using Liverpool Malaria Model (LMM) driven by System 4. Skill is demonstrated over Botswana, particularly for forecasts issued in November, validating against laboratory confirmed cases of malaria. This is an improvement on previous work where the LMM was driven with the DEMETER seasonal hindcasts. Where malaria data is not available, System 4-driven LMM hindcasts are compared to LMM driven by ERA-Interim in a tier-2 validation context. Skill is demonstrated at the epidemic fringe of the Sahel and in north west Malawi, whilst the Gulf of Guinea shows no skill. This is consistent with previous work suggesting the LMM performs better in epidemic than in endemic regions. A method

for interpreting hindcast validation results as uncertainty quantification is then presented.

Finally, the uncertainty in the relationship between seasonal average climate and malaria risk is analysed, using the LMM driven by the 20th century reanalysis dataset. The relationship parameters describing seasonal average climate and malaria risk is explored and impact surfaces are created, relating seasonal average temperature and precipitation to average seasonal malaria incidence. The robustness of these impact surfaces is investigated by comparing the surfaces associated with different LMM survival schemes. A method of combining impact surfaces based on tercile categories is described and implemented and it is demonstrated how the resulting graphic could be integrated with a seasonal ensemble forecast system. Such a tool is potentially useful for decision-makers, allowing an intuitive visual communication of the quantified uncertainty in predicting climate-driven disease risk at seasonal timescales.

Acknowledgements

Thanks first go to Andy Morse for facilitating my work, connecting me to others in the field and giving me the space to explore my own ideas. I also acknowledge NERC for funding my studies.

I'd like to express gratitude to all in the Roxby building, for providing a welcoming and friendly working environment during my three years in Liverpool. Thanks particularly to the denizens of room 111: Cyril, whose expertise helped make the analysis what it is and whose unique way of interfacing with his computer improved my French and my ability to focus, Andy Heath, whose presence helped make the office such a great place to work and kept me sane and finally to Anne for interesting and useful discussions of work, particularly regarding malaria modelling;. Further appreciation goes to all of the above for proof reading and help with figures.

Outside of work, my thanks goes to everybody at and connected with Huski house; you've made my time at Liverpool. Not forgetting Oliver at the treehouse, we've had some great times there and I'm sure there'll be more. Finally thanks to friends outside Liverpool and my family, for being a bedrock of certainty, and to Megan, whose love and support always keeps me going.

Contents

CHAPTER 1

Motivation

Climate and disease are interconnected; anomalous conditions of temperature and precipitation can cause disease outcomes different from normal. For example, low seasonal rainfall over a region can reduce the local population of mosquitoes, reducing the rate of malaria infection and the disease burden. However the disease landscape is highly non-linear; temperature and precipitation interacts in many ways with the life-cycles of disease vectors and pathogens. This prevents simple models of cause and effect and so to accurately predict the effects of the climate on disease risk is not a simple task.

Climate change is happening (IPCC, 2007). Global temperatures are expected to rise over the coming century, masking regional variations in the magnitude of the increase, whilst trends in regional precipitation are less certain. In any case it is likely that in many regions the environmental conditions affecting disease risk will change. This has serious risk implications for not only the many societies without adequate health infrastructure to act as a buffer against life-threatening diseases, but also for more developed societies; changes in local climate may allow diseases to emerge in regions where previously the environment was unsuitable. Climate-driven epidemiological models can be used to analyse these risks, and potentially to provide forecasts.

Epidemiological models are generally of two types, dynamical and statistical. Dynamical models directly simulate disease dynamics, whilst statistical models are based on empirical relationships between climate and disease variables estimated from lab and field work. Both kinds of model can be used to make predictions of climate-forced changes in disease when linked with climate models. The term 'climate model' can be used to describe anything from a simple energy balance equilibrium model 'simulated' with a pen and paper through to earth simulator models which require the computing power of the biggest supercomputers in the world.

The most complex climate models are based on general circulation models (GCMs). These explicitly simulate the large-scale dynamics and thermodynamics of the atmosphere and ocean, using physically-based equations. They can be used to make predictions of the future state of the climate and have different targets: seasonal prediction aims at the next few seasons, decadal prediction at the next decade, whilst when used in climate prediction mode they simulate up to a century ahead and further. This thesis is based primarily on using seasonal and decadal GCMs initialised with observations, considering the uncertainty present when they are used to drive disease models and make predictions.

If these predictions are ever communicated to potential users, it vital that they are accompanied by an estimation of confidence. A forecast with an uncertainty estimate is more useful than one without, even if the uncertainty is large as it has been shown that people make better decisions when forecast uncertainty is communicated (Joslyn et al., 2007).

How then to represent the uncertainty of a model model prediction? If possible, the most useful way is to quantify the uncertainty in forecasts; that is, to provide error bars, or a range of outcomes into which there is a high belief the future will fall. However, an initial step (prior to quantification of uncertainty) is to validate the model; to compare forecasts against 'what really happened'. If the model does not perform better than chance then it is said to have no skill. If a predictive model has no skill then there is no practical use in quantifying forecast uncertainty, in the same way that there is no use in giving a confidence range on a prediction for the result of tonight's lottery result. If however the model does have skill, work on estimating and quantifying uncertainty can begin.

There are many places from which uncertainty can arise in the world, and there is no one way to quantify it; many methods to do so have been described (for a good review see Halpern, 2003). Communication of uncertainty is also an important step and should be considered; without proper communication, the job is half done. It is important to consider one's audience since the message about uncertainty must be tailored to fit. People with different backgrounds may interpret the same information in different ways, as they use different language and have different priorities; a mathematician is likely understand the statement 'the model prediction is uncertain' in a very different way to a policy maker might. Thus communicating uncertainty to non-specialists requires a certain subtlety.

In this thesis work is presented on the topic of the quantification of uncertainty in climate-driven disease risk. Chapters two and three contain a literature review and methodology, and following this the results chapters are divided into two parts. Part I deals with climate forecasts at decadal and seasonal timescales, beginning with chapter

4, where the first multi-model decadal hindcast dataset produced as part of the ENSEMBLES project (Van Der Linden and Mitchell, 2009) is analysed. Chapter 5 then considers how forecasts of seasonal climate over Africa and India have evolved over the past decade, comparing the hindcasts produced by seasonal climate models used in the DEMETER (Palmer et al., 2004 and ENSEMBLES projects with the skill of System 4 (a state-of-the-art seasonal climate model, developed by and run at the European Centre for Medium-Range Weather Forecasts). Chapter 6 then considers the forecasts offered by System 4 in more detail.

Part II considers the uncertainty related to disease predictions at seasonal timescales, focusing on malaria. The Liverpool Malaria Model (Hoshen and Morse, 2004) is employed, and chapter 7 describes the quality and uncertainty of forecasts made when it is driven by the System 4 seasonal forecasts. Chapter 8 then relates the use of the LMM to study the link between seasonal average climate conditions and malaria risk, and considers the quantification of uncertainty related to the malaria model. Finally, chapter 9 contains a summary of main conclusions and a discussion.

The quantification of uncertainty in forecasts of climate-driven disease risk is a broad interdisciplinary topic; in this thesis the following questions are addressed:

- At what timescales are climate predictions good enough to provide useful decision-relevant information about future disease risk? (chapters 4, 5, 6 and 7)
- Can decadal climate models make useful disease predictions? (chapter 4)
- What is the quality of forecasts from state-of-the-art seasonal climate models and where is it sufficiently high to drive to disease models? (chapters 5 and 6)
- How good are climate-based predictions of malaria at seasonal timescales, and what is the associated uncertainty? (chapter 7)
- Can seasonal averages of temperature and precipitation be used to predict disease risk? (chapter 8)

The following are a list of unique research contributions made by this thesis:

- Validation of annual and seasonal temperature and precipitation anomalies and trends at the decadal timescale, at multiple spatial scales, using the ENSEMBLES decadal hindcasts (chapter 4)
- Comparison of temperature and precipitation predictions over Africa and India from DEMETER, ENSEMBLES and System 4, charting the progression of seasonal climate prediction skill over the past decade (chapter 5)
- Evaluation of how the skill of System 4 forecasts varies with lead time (chapter 6)

- Evaluation of the skill of coupled System 4-LMM forecasts at seasonal timescales over African regions (chapter 7)
- The creation of a prototype decision-support tool for malaria, linking predictions of average seasonal climate with malaria risk and quantifying malaria model uncertainty (chapter 8)

CHAPTER 2

Literature review

This chapter reviews the literature covering the quantification of uncertainty in climate driven-disease risk. It is divided into two halves. The first covers the forecasting of climate-driven disease, starting with the interaction between climate and disease, before covering climate models and seasonal to decadal predictability. The section finishes with a discussion regarding the use of climate-driven disease models to make early warnings of disease risk.

The second half of the chapter considers uncertainty, starting from a fundamental review of what it means to quantify uncertainty and with a look at the methods used to do so in climate models. The possibility of defining a generalised uncertainty framework follows and the section concludes with a short section on the considerations necessary for effective communication of uncertainty.

2.1 Forecasting climate-driven disease risk

2.1.1 The interaction of weather and climate with disease

The distinction between weather and climate is one of averaging; climate is the long term average of the weather, normally taken to be 30 years. Climate has spatial and temporal variability; spatially, climate varies with latitude, altitude, land surface type and topography, with each regional climate determined by its unique location and surroundings. Temporally, weather varies on all time-scales, with a dependence on location; at the equator the highest temporal variability in temperatures is the diurnal cycle as the earth turns first toward and subsequently away from the sun, and at higher latitudes the highest source of temporal variability is the seasonal cycle, as summer turns to winter and back again. Rainfall exhibits a much higher variability than temperature and depends strongly on regional characteristics, defying classification by simple rules (Gray, 2007).

The weather is organised into systems which interact on a wide range of temporal and spatial scales. These range from short-lived convective systems which affect kilometre-scale regions over a few hours, to low-frequency oceanic modes oscillating on centennial timescales and acting over wide regions of the globe. At an intermediate scale, the large-scale monsoon systems providing rainfall to the Indian subcontinent and West Africa are created by differential patterns of heating of the ocean and land. Finally at smaller scales regional rainfall patterns vary with proximity to the sea, altitude and topography (Gray, 2007).

Regional characteristics and variability of weather and climate are important forcings on disease; a disease will invariably arise within a permissive climate where a competent host and vector population intersect (Reisen, 2010). That is, for a specific disease there is a limited range of climatic conditions - the climate envelope - within which the corresponding pathogen, vector and host species can survive and reproduce.

Introducing some epidemiological terms:

- Pathogen: a microbe or micro-organism which causes disease in its animal host
- Host: the animal in which a pathogen lives
- Vector: any agent which carries and transmits an infectious pathogen into a living organism (usually a mosquito or tick)
- Vector-borne disease: a disease which transmitted via a vector to humans or other animals

In general the meteorological variables most strongly linked to disease are temperature, precipitation and humidity. Other factors such as wind and sunlight duration can also be important (Reiter, 2001; Rodó et al., 2011; Shea et al., 2008; Sultan et al., 2005). To a first approximation, temperature governs the rates at which biological processes occur, though the interactions are non-linear. Some general effects of and links between temperature on vectors, pathogens and hosts are (adapted from Gubler et al., 2001):

- Higher temperatures increase the daily mortality of some vectors e.g. *Culex Tarsalis* (Reeves et al., 1994)
- Higher temperatures reduce the incubation period of the pathogen in some vectors, e.g. Western Equine Encephalomyelitis and St. Louis Encephalitis in *Culex Tarsalis* (Reisen et al., 1995)
- Increases in temperature changes vector distribution; generally polewards and upwards towards higher elevations, e.g. *Ixodes ricinus*, the vector for Lyme disease and Tick-Borne Encephalitis (Mills et al., 2010)
- Warmer temperatures may increase the rate at which mosquitoes bite hosts, e.g. the female *Anopheles* mosquito, the vector for malaria (Githeko et al., 2000)

- Temperature may affect the length of the transmission season of diseases (Hardy et al., 1990)

Along with temperature, rainfall governs vegetative structure, providing a habitat for vectors. The spatial and temporal characteristics of rainfall generally govern the extent of the areas for possible disease transmission (Reisen, 2010). Listed below are the main effects of precipitation on vectors, pathogens and hosts (from Gubler et al., 2001):

- Vector numbers may increase after heavy rainfall events,e.g. *Aedes* mosquitoes, the vector for Rift Valley fever (Anyamba et al., 2001; Mondet et al., 2005)
- Humidity (and therefore precipitation) influences the mosquito biting-laying cycle and mortality (Detinova, 1962)
- Excess rain or snowpack can eliminate habitat by flooding, decreasing vector population (Gubler et al., 2001)
- Low rainfall can create vector habitats by causing rivers to dry into pools (Gubler et al., 2001)
- There are few direct precipitation effects on pathogens but there is some evidence that humidity affects malarial parasite development in the *Anopheles* mosquito host (Gubler et al., 2001)
- Increased rain can increase vegetation, food availability, and therefore host population size (Gubler et al., 2001)
- Increased rain can cause flooding: decreasing host population size but simultaneously increasing human contact (Gubler et al., 2001)

Quantifying these relationships between climate and disease can potentially facilitate the creation of early disease risk warnings, if forecasts of climate anomalies can be made. The following section describes the tools, methods and theory underlying the climate models used to forecast climate.

2.1.2 Climate models and climate modelling

In general, climate models are used to better understand the climate system. Some models can be used to make short term forecasts, whilst others are used to make long-term projections of future climate (see McGuffie and Henderson-Sellers, 2005 for a thorough description of climate models, and Taylor et al., 2012 for a description of all the streams climate modelling experiments used in the most recent phase of the Climate Model Intercomparison Project).

There is a subtle distinction between forecasts and projections which should be made here. A forecast is an prediction of what is actually expected to happen in the future, whereas a projection describes a hypothetical 'what-if' scenario, made with certain

assumptions (IPCC, 2007). This thesis does not consider projections, instead is concerned with short term climate forecasts. In particular forecasts are consider for seasonal to decadal time scales, where validation is possible and usefulness is arguably greater (Washington et al., 2006).

Nearly all climate models can be described as either statistical or dynamical models. Statistical, or empirical, models consist of relationships between macroscopic climatic variables which have been derived from past observations. To make a gross simplification, they forecast the future by extrapolating the statistical relationships observed in the historical record. In reality they are more sophisticated than this (see Dool, 2007 for an excellent overview). In this thesis however statistical climate models are not considered, the focus is instead on dynamical models.

Arguably dynamical models have an advantage over statistical models, in that they have more potential for improvement. With access to a set of predictors and predictand, a well-trained forecasting statistician could potentially produce the best possible statistical model in a fairly short time. Aside from getting new observations, or finding new predictors, more effort is unlikely to bear fruit. On the other hand, a dynamical climate model has many things which can be improved: addition of models for systems not previously considered (for example including the land surface, or the carbon or nitrogen cycles), higher resolution with increasing computer power, and better parametrisations. Dynamical models also allow researchers to test hypotheses, for example by looking at the impact of the impact of the shut-down of the thermohaline circulation on global climate (Vellinga and Wood, 2007). A review of dynamical climate models follows.

Dynamical Climate Models

Dynamical climate models, or general circulation models (GCMs) evolved from numerical weather forecasting. The first instance of this occurred with Lewis Fry Richardson's pioneering six-hour forecast for 20th May 1910. The story of this forecast and the full history of climate modelling has been described in interesting detail elsewhere (Lynch, 2008) but it is useful to summarise here as a platform from which to describe GCMs.

Richardson's forecasting method began with a division of the forecast spatial target region into a grid. Each grid box was then given values for various meteorological parameters corresponding with observations made at a certain point in time, creating an initial state. The calculation of the expected movement of mass and heat in a small time between neighbouring boxes was made, using the equations governing the

movement of fluids (i.e. the Navier Stokes equations), plus thermodynamic equations for heat transfer. This was repeated for many successive time steps, producing the worlds first numerical weather forecast. This principle underlies numerical weather prediction and dynamical climate models today, the main difference being that the calculations are now done on supercomputers rather than with a pencil and paper. A schematic of a GCM is given in figure 2.1.

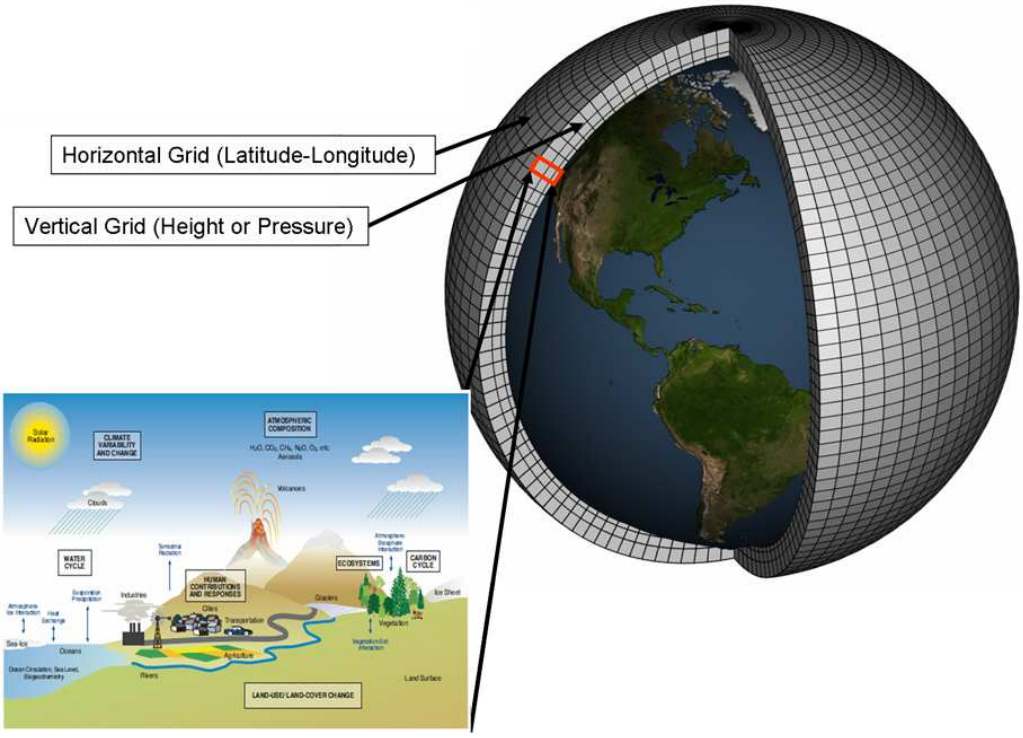


Figure 2.1: A schematic of a global climate model (adapted from GFDL, 2013).

Advances in computing technology have enabled numerical weather prediction models to become the major tool used to forecast weather and climate. Since Fry's original forecast for 1910, numerical weather forecasts have become routine around the globe, with the first real-time forecast issued in 1954 (Harper et al., 2007). The resolution and complexity of climate models has also increased; models used in the Intergovernmental Panel on Climate Change (IPCC) First Assessment Report in 2001 had a horizontal resolution of 500km, with 10 vertical layers in the atmosphere and a slab ocean, whilst those used in the Fourth Assessment Report in 2007 typically used 110km horizontal resolution with 30 layers each in the ocean and atmosphere (IPCC, 2007). Models also now have more complexity; most models used for forecasting today have at least atmosphere and ocean components, with the most complex models (earth system models) including a full three dimensional coupled atmosphere and ocean, an interactive cryosphere (ice sheets and shelves), along with dynamic simulation of the

carbon cycle, the nitrogen cycle and the biosphere (McGuffie and Henderson-Sellers, 2005).

There are many challenges related to running a GCM, though two of the major ones are initialisation and parametrisation. Initialisation relates to the starting point of any forecast, the initial state of the system. In general the closer to the initial reality the model is, the better the forecast. Since it is impossible in reality to replicate with exact precision the state of the atmosphere at an instant in time, mathematical techniques are employed in order to take what is available - sparse observations from a multitude of sources - and transform them to a 'best guess' of a smoothly varying field representing a snapshot of the climate which can be used as an initial state for the model grid. These techniques are known as data assimilation, and their implementation is vital to the success of a forecast. For a description of data assimilation, see Wang et al., 2000.

The need for parametrisation arises due to the limited size of model grid cells. Whilst cells are generally of sufficient size to capture the large-scale flow dynamics, processes which occur on scales smaller than the grid are essentially invisible to the model (Stensrud, 2009). These processes are known appropriately as sub-grid-scale processes and they are related to important aspects of the climate, such as convection, land surface processes and cloud cover. To include these unresolved processes models use parametrisations, additions to the large scale equations which attempt to represent the bulk effects of sub grid-scale processes. These parametrisations contain coefficients and values which are often difficult to determine experimentally, as such they can be significant sources of uncertainty in GCM forecasts. Research on parametrisation is an active field.

Greater understanding of the predictable components of the atmosphere along with higher resolution of models and better initial observations has allowed time-scales of prediction to expand. Medium range (up to 30 days ahead) and seasonal forecasting (1 month-1 year) have been developed in the past few decades and most recently decadal modelling (over 10 years) has emerged. In this thesis focus is on the potential to drive disease models with seasonal and decadal models. The next section contains a short review of seasonal to decadal predictability.

Seasonal to decadal predictability

Edward Lorenz demonstrated how unavoidable uncertainties in initial conditions will invariably grow and contaminate a weather forecast (Lorenz, 1963). This sensitivity to initial conditions (sometimes referred to as the 'butterfly effect') limits the time period over which even a perfect model can yield skilful weather forecasts to about two weeks.

However at longer lead times there is nonetheless some skill in predicting anomalies in the seasonal average of the weather i.e. anomalies of the climate.

Most skill in predicting seasonal climate comes from the slowly changing conditions at the earth's surface. The most important surface condition affecting climate is the sea surface temperature (SST), especially SST in the tropics, whilst soil moisture and snow cover also offer some predictability (Palmer and Hagedorn, 2006b). For SST, predictability comes from the thermal inertial of the oceans; when ocean temperatures are higher than normal, they usually remains that way for several months. Sometimes this can be for as long as a year or more, such as during the El Niño or La Nina episodes of the tropical Pacific SST. Similarly, when there is high soil wetness or snow cover, it often takes at least several weeks to return to normal - each only a limited portion of the excess evaporates or melts.

Seasonal forecasts have a level of accuracy that whilst not perfect is above the level of chance, and seasonal predictability varies by location; in the tropics and near the coast it is generally higher, with inland areas and mid-latitude zones offering less potential for prediction at seasonal timescales (Palmer and Hagedorn, 2006b). However in some regions it is good enough to make a difference on sectors where climate variability impacts society, such as agriculture, energy production and health (Tall et al., 2012).

Information on decadal timescales would be also be useful, particularly for climate change adaptation, as it is a key planning horizon for governments, business and other societal entities (Cane, 2010). Evidence from idealised predictability studies and initialised decadal climate projections suggests that some aspects of climate variability may be predictable for a decade or longer in advance, which has prompted further research into the potential for decadal climate forecasts (e.g. Keenlyside and Ba, 2010; Meehl et al., 2009; Mehta et al., 2011; Murphy et al., 2010; Oldenborgh et al., 2012).

On decadal scales, predictability is believed from low-frequency climate modes such as the Atlantic Multidecadal Oscillation, a basin-wide fluctuation of sea surface temperatures in the North Atlantic with a period of around 70 years (Schlesinger and Ramankutty, 1994). It may also come from the Pacific Decadal Oscillation (a low-frequency mode of Pacific climate variability, Mantua and Hare, 2002), as well as from land and sea ice (Murphy et al., 2010). It has been shown that initialization of decadal climate models can improve skill of climate simulations (Pohlmann et al., 2009).

Predictability may also arise from the ability to predict the evolution of external boundary condition forcings; namely changes in greenhouse gas concentrations (Keenlyside and Ba, 2010). Boundary condition predictability underlies the way in which climate change projections are made, where the forced climate response can be

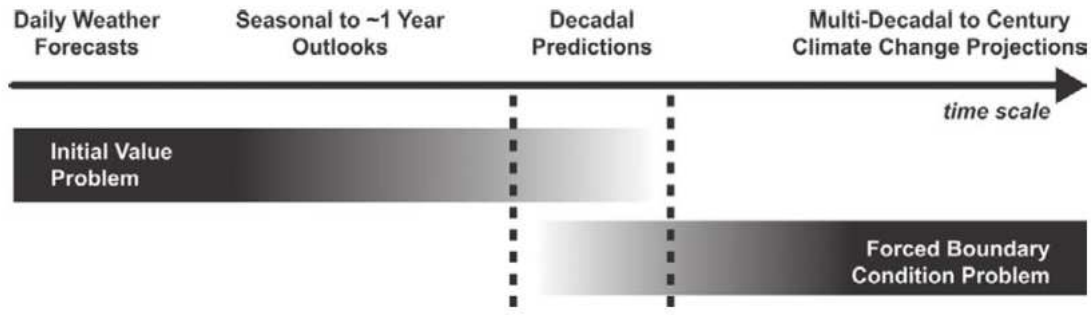


Figure 2.2: Timescales of climate modelling. Decadal predictions are potentially both an initial value and a forced boundary condition problem (from Meehl et al., 2009).

estimated from projected rises in greenhouse gases concentrations, comparing to 'pre-industrial' runs for which greenhouse gas emissions are kept constant. This interplay of initial and boundary conditions is a unique aspect of the decadal prediction problem; whilst seasonal forecasts take their skill from initialization of variables such as SST, and climate change projections are based entirely on boundary conditions such as changes in greenhouse gases and radiative forcings, decadal predictability potentially comes from both. This is illustrated in figure 2.2.

The extent to which variations in decadal climate are predictable on the time and space scales to predict climate impacts is an open question, and one which is dealt with in this thesis. On spatial scales smaller than the subcontinental scale, it takes several decades for the forced temperature signal to emerge (Karoly and Wu, 2005; Knutson et al., 1999). The situation becomes more difficult for other climate variables, such as precipitation, where even large-scale forced changes are only marginally separable from internal climate variability (Min et al., 2008; Zhang et al., 2007). Thus, some unresolved questions remain regarding not just how decadal predictions should be conducted, but also regarding the quality and usefulness of results at the timescales which current decadal prediction experiments are targeting.

To systematically look the skill in decadal climate predictions, the first multi-model decadal hindcast set was produced, as part of the ENSEMBLES project (Hewitt and Griggs, 2004). This hindcast set consists of state-of-the-art models from the major meteorological centres in Europe, run in decadal prediction mode and making 10 year forecasts. It is this dataset which is studied in this thesis, considering whether the skill of the hindcasts is sufficient to drive a disease model. Further details of the models and results are presented in chapter 4.

Though decadal climate prediction is still an experiment, seasonal climate models are

currently used routinely to make useful forecasts of seasonal climate. This has motivated attempts to develop climate based models for seasonal disease forecasting. A short review follows.

2.1.3 Using climate-driven disease models for early warnings

The purpose of climate-driven disease models is to provide early warning of impending epidemics based on climate information, forming the basis of health early warning systems. Such systems would be invaluable for epidemic preparedness and prevention, though using climate information to make disease forecasts is not new. The use of climate data for predicting outbreaks of infectious diseases dates back to 1923, to work by Gill and others in India, who developed an early warning system for malaria based primarily on rainfall conditions (Gill, 1923). The model itself was used to predict epidemics from 1921-1942 in 29 districts of the Punjab and formal assessment of the model's performance indicated that its accuracy was significantly better than would have been obtained by chance (Swaroop, 1949).

Since that time, our modelling ability has improved both conceptually and computationally, with the development of predictive models for disease which can be linked to climate forecasting models. These fall into two categories, process-based models and statistical models.

Process-based disease models are also known as mechanistic or biological models. They directly represent processes occurring in nature using mathematical relationships derived from observation (laboratory and field work) and theory. Most process-based models of infectious disease are for vector-transmitted diseases, and they attempt to simulate processes occurring in the host-pathogen-vector complex under varying assumptions. As an input, they use mathematical relationships between climate variables (usually temperature and precipitation) and the rates relating to vector and pathogen life-cycles[1] (Lafferty, 2009).

Process-based models also contain parametrisations which require knowledge of relationships between climate variability and vector/pathogen life-cycle rates, which are sometimes difficult to determine experimentally. The models in general only project the potential for transmission rather than actual transmission because the potential niche is always larger than the realised niche (Lafferty, 2009). Put simply, temperature and precipitation cycles create environmental limits for development of pathogen and vector populations and the amount of disease actually present in a population is

[1] Examples of these rates are the larval development rate, mosquito biting rates and the daily mortality rate, among others.

constrained by other predominant factors (such as planned interventions in the environment and health care systems). Barriers to dispersal and biotic interactions can also exclude species from parts of the potential niche (Lafferty, 2009).

Various models for have been used to study the possible effects of climate change on diseases (Bhattacharya et al., 2006; Martens et al., 1999; Peterson, 2003; Peterson and Shaw, 2003; Tanser et al., 2003; Thomas et al., 2004; Urashima et al., 2003). However few have been linked directly with climate models making forecasts on shorter timescales, with some notable exceptions where a seasonal ensemble climate prediction system has been linked with a dynamic process based model for malaria (Ermert et al., 2012; Jones and Morse, 2012; Jones and Morse, 2010; Jones, 2007; Morse et al., 2005). The dynamical malaria model used in these studies is the Liverpool Malaria model (LMM, Hoshen and Morse, 2004), and is used within this thesis. Details of this model are given later, in section 7.1.1.

The other main group of climate-driven disease models are statistical models. They relate observations of climate to disease observations and do not explicitly simulate processes. These are not studied in this thesis and are not discussed further here.

Often, non-climatic variables are not included in climate-disease models. Climate is only one of several important factors influencing the incidence of infectious diseases. Other important things to consider include socio-demographic influences, such as human migration, poverty, transportation, availability of health services, drug resistance and nutrition; as well as environmental influences such as deforestation, agricultural development, water projects and urbanization (McMichael et al., 2006). The absence of these factors in disease models is a major source of uncertainty.

Despite this, a climate-driven disease model can still be useful, providing a backdrop on which to consider the interaction of other factors. As has been discussed in this section, the climate has links with disease, and has some predictability, allowing the creation of climate-driven models and potentially early warning systems. However, a forecast without an estimate of uncertainty is most likely misleading and potentially dangerous. As Hendrik Tennekes put it:

> *No forecast is complete without an estimation of forecast skill (Tennekes, 1992).*

It is important to be clear about underlying assumptions and to quantify the uncertainty in predictions. Uncertainty is the subject of the second half of this literature review.

2.2 Uncertainty

A statement of uncertainty is a way of expressing a level of belief or confidence. Uncertainty will always be with us - there is a long philosophical tradition which has shown that even our most firmly held beliefs can have doubt cast upon them (Descartes, 1641; Najm, 1966). To borrow part of a famous quote from George Box;

> *All models are wrong* (Box, 1979).

The word model stands here for any construct which aids the understanding of a real-world system. It can represent not just the mathematical models dealt with in this thesis, but also conceptual models and mental models. There is always a mismatch between a model and reality and uncertainty is always present.

Before continuing with discussion of uncertainty, it should be separated from risk. Though different concepts, often the words are lazily used interchangeably. Simply defined: uncertainty is associated with a belief about the reality of a statement or idea, whilst risk is associated with a potential event in the world. Risk can be defined as the likelihood of an event multiplied by the magnitude of its potential impact. There is uncertainty associated with both the likelihood and with the magnitude of impact, which can increase the perceived risk by increasing the range of potential harm. Because of this, higher uncertainty is generally associated with higher risk. For example, the forecast for tomorrow's weather may be much more uncertain than the usual prediction of fine weather, meaning a decision to go rock climbing is risky one. However, if the forecast is for torrential rain with high confidence the decision to go climbing is still a risky one. That is, risk can be associated with low or high uncertainty and despite their close relationship, risk and uncertainty are separate concepts.

Considering forecasts then; the future is inherently uncertain and any model which attempts to predict it is bound to uncertainty. In general, the further into the future one attempts to predict, the more the uncertainty grows. An example of this is our ability to say with more certainty what we will be doing tomorrow than what we will be doing next year. This does not prevent making long-term predictions, though it can guide us to understand which predictions are more trustworthy and useful than others. Our level of certainty in a prediction of the behaviour of a system depends on its structure and tendencies and our knowledge of these. It also is related to the capacity for one system to be affected by others. So whilst it is true that there is always uncertainty present when we make a prediction, this does not prevent us doing so (as we do every day of our lives). To complete George Box's quote then;

> *All models are wrong, but some are useful.*

Forecasts from even the simplest models have associated uncertainty, but as long as we have some idea of its extent, models can be useful. As it turns out however, people are not particularly good about thinking about uncertainty (Gigerenzer, 2003), necessitating its consideration in an objective way.

Though the easiest thing to do would be to ignore uncertainty, this causes end-users of forecasts to act as if predictions are certain and potentially leads to bad decisions. A more responsible way to address uncertainty is to qualitatively acknowledge the possibility that a future projection may turn out to be wrong, in known and unknown ways. This may be the only possible way to include some types of uncertainty in results; at the very least the assumptions upon which a prediction rests should be made clear.

Another way to deal with uncertainty is by presenting output in the form of alternative versions of the future, using 'what-if' scenarios, or projections. This is useful if several possibilities for the future can be described, however assigning probabilities to each may be difficult. The scenario approach has been used by the IPCC, where discrete projections are given for different emissions scenarios (IPCC, 2007). This method is useful not just as a way for accounting for uncertainty in knowledge in future emissions, but also as a device to illustrate the potential effects of different choices, clearly showing effects of different courses of action. It can be used to effectively communicate 'best' and 'worst' case scenarios.

A final way to deal with uncertainty in predictions is to address it systematically, by quantifying it. This method is generally to be preferred if it is possible to implement; well-quantified uncertainty allows end-users to make the best decisions (Joslyn et al., 2007). However the ease by which uncertainty can be quantified depends on the context of the situation. A discussion of common methods used to quantify uncertainty follows.

2.2.1 Quantifying uncertainty

On the question of how to quantify uncertainty, a useful discussion of the most common methods is found in Halpern, 2003. The idea of 'possible worlds' is introduced, that is, the worlds considered possible given our current understanding. The span of possibilities of these worlds can then be seen as a qualitative measure of uncertainty. Fewer possibilities mean there is less uncertainty, whilst conversely more possibilities suggest more uncertainty as to the truth. All formal methods to quantify uncertainty, of which there are several, essentially start from this point.

Probabilistic methods are by far the most well-known and well-used methods, but they are not the only ones. Other numerically-based methods include Dempster-Shafer belief

functions, possibility measures, and ranking functions. There are also non-numerical methods such as using relative likelihood or plausibility measures. Descriptions of the theory behind these alternative methods is given extensively in Halpern, however they are not used in this thesis and are not discussed further.

The advantage of using probability over all other methods is that it is well understood - organisations relevant to climate predictions (e.g. government agencies, insurance companies) are generally used to working with probabilistic information over the alternative methods listed above. A number of arguments also suggest that in certain situations it is the only 'rational' way to represent uncertainty (Halpern, 2003).

There are issues with probability; when numbers are not available or when one is not prepared to assign probabilities to different events other methods may be more appropriate. Furthermore it is not good at representing ignorance. Despite these issues, probability is an appropriate tool to use to represent uncertainty in computer modelling, as numbers are ubiquitous and the nature of simulation makes it easy to calculate probabilities and the range of outcomes from model output. Computer modelling essentially repeatedly simulates all of the possible worlds consistent with our understanding, the output from the model gives an estimation of the uncertainty in the model world and the fraction of simulations in which an event occurs can be equated to its probability of occurrence.

It is important however to always keep in mind that the model world is not and never will be identical to reality and that an event in the model world may not be identical to an event in reality. Therefore there is always some uncertainty associated with computer model output, but arguably no more than that which springs from the predictions we make every day using our mental models. One important difference is that computer models enable uncertainty ranges to be estimated numerically. It is also much easier to be clear about the assumptions a prediction rests upon when the model is written down explicitly. Finally, systematic validation of the performance of a computer model is possible and is generally performed as standard, depending on the discipline. Mental models on the other hand are rarely subject to the same kind of validation.

Quantification of uncertainty has been considered in different modelling disciplines to varying degrees. The following section describes uncertainty and its quantification in climate modelling, followed by a discussion of uncertainty in climate-disease modelling and finally a description of a general uncertainty typology.

Quantifying climate modelling uncertainty

Climate modelling prediction uncertainty arises from three main sources: uncertainty in initial conditions, scenario uncertainty, and the uncertainty intrinsic in models; model uncertainty (Collins, 2007).

The climate is a chaotic dynamical system and slightly different initial states can evolve into considerably different states in the future. Climate models are initialised with observations as initial conditions, however, these cannot be known with exact precision due to measurement and sampling limitations. For this reason the inevitable uncertainty in initial conditions leads to a model forecast which diverges from reality after a short time. Initial condition uncertainty is important for short range forecasting but becomes less important as predictions are made further into the future as errors saturate.

Scenario uncertainty pertains generally to climate models simulating end of century climate, and relates to the large variation in different emission pathways and associated radiative forcing changes. At shorter time-scales the difference between scenarios is negligible, it is only when integrated over multi-decadal time-scales that the scenarios diverge significantly.

Model uncertainty arises from the fact that no model contains a perfect representation of the climate system. This incompleteness and inadequacy of models comes from a variety of sources: formulation of model equations, uncertainty in parameter choice, limited representation of processes due to limited resolution or from an incomplete knowledge of processes. Non-linear interactions between components of the climate also contribute to this uncertainty. Model uncertainty affects predictions made at all time-scales, and can be further sub-divided into parameter uncertainty (uncertainty in the parameters that control parametrised physical processes) and structural uncertainty (uncertainties in choices made when coding the resolved processes).

Scenario uncertainty can be dealt with as described previously, by integrating models with different scenarios and presenting the outcome as a set of discrete decisions. Initial condition uncertainty is harder to deal with, but one way of quantifying this error is to employ ensemble forecast systems, developed originally for numerical weather prediction. An ensemble forecast is a collection (ensemble) of forecasts where each member of the ensemble is initialised with different set of initial conditions consistent with observations. This allows uncertainty in initial conditions to be quantified; a forecast can be produced in the form of a probability distribution function (PDF) weighted according to the support for a prediction by individual ensemble members. Ensembles also give an idea of the predictability of the system; if the PDF for a forecast

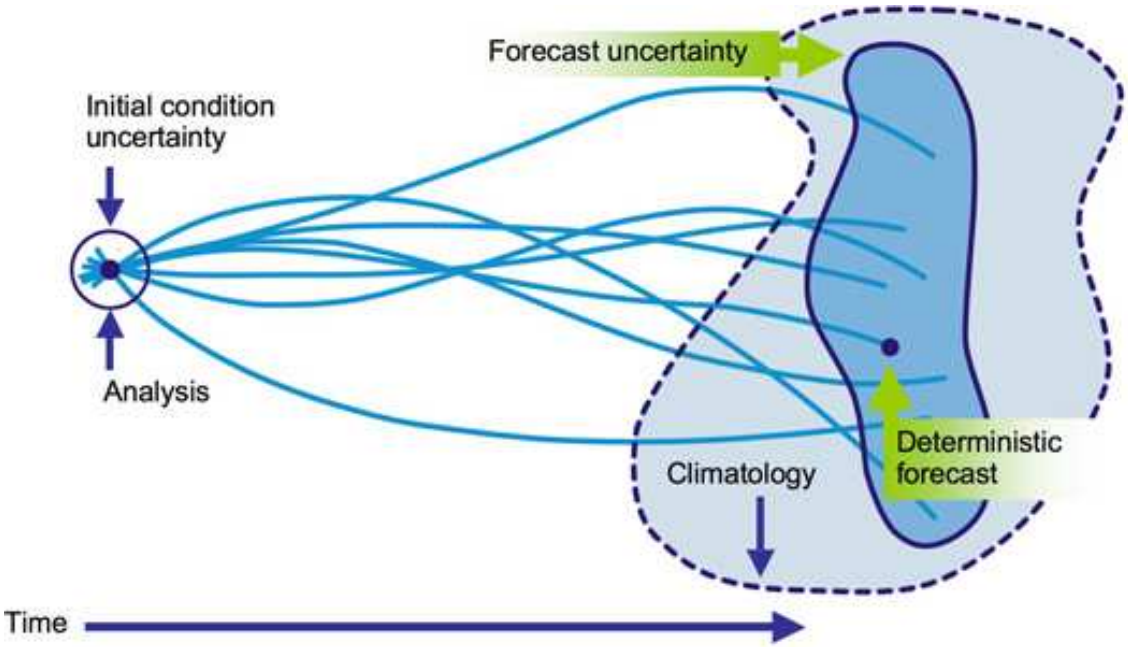


Figure 2.3: A visualisation of how using an ensemble of initial conditions quantifies initial condition uncertainty (UKMO, 2013). Multiple initial states consistent with observations are used to run a model; the spread of the output gives a measure of the effect of this initial condition uncertainty and provides a probabilistic forecast. Compare this to a deterministic forecast which has no estimation of this uncertainty.

variable is spread over a wide domain this indicates unpredictability and uncertainty as there is no general consensus among ensemble members. Conversely if the PDF is narrow this indicates high predictability as a high proportion of the 'possible worlds' simulated have the same outcome[2]. This is shown in figure 2.3.

Ideally, to quantify well the uncertainty relating to initial conditions an ensemble would be created containing a member for every possible permutation of the atmosphere and ocean consistent with observations. However the number of degrees of freedom in which to choose perturbations vastly exceeds the largest practical size of an ensemble. Instead compromises are made so that the maximum amount of possible variance in initial conditions is captured by the smallest possible size of ensembles (Palmer, 2000). This suggests that probability distributions calculated by initial condition ensemble prediction systems do not fully capture the range of initial condition uncertainty, however they are a pragmatic solution to the problem of quantifying uncertainty considering computing constraints. As such, PDFs from an ensemble forecast should

[2]Note that a narrow ensemble spread in a model only indicates high *potential* predictability, it does not necessarily imply forecast accuracy. A narrow spread from a model which historically has never made an accurate forecast is analogous to meeting a man in the street who is convinced the world will end tomorrow. Certainty is not the same as skill; precision does not equal accuracy.

not be treated as absolutely representing initial condition uncertainty, but instead as capturing a portion of the uncertainty - there is still chance that the true future could lie outside the range of the spread. It could be argued however that whilst such systems do not completely quantify the uncertainty associated with a prediction, they are able to determine the uncertainty associated with one prediction as being more or less uncertain than another.

Considering model uncertainty, there are several approaches to quantifying it. One is the multi-model approach, which collates output from several different models created and run from different modelling centres. Each model is run as an individual initial condition ensemble to quantify initial uncertainty, but together they also aim to quantify model uncertainty and are known as a multi-model ensemble. The idea behind this is that since each model is structurally different, allowing some exploration of 'model space'. The multi-model approach has the advantage in that it addresses structural uncertainty; different models in the ensemble will have a different structure. The drawback with this approach is that it is not objective - models are not developed independently, developers will have made models under similar influences and will have influenced each other in formulation of the models. It is therefore unlikely that model output is only a subsection of the complete model space (Collins, 2007).

One way to deal with parameter uncertainty is the perturbed physics approach (e.g. Murphy et al., 2007). In this method models are integrated repeatedly with different choices of parameters. Whilst this allows some quantification of parameter uncertainty, a drawback are that the parameter space is huge and sampling the whole space is computationally costly and so only a sub-sampling is possible. This is potentially an issue due to the non-linear nature of the interaction between parameters. Furthermore, perturbed physics approaches do not deal with structural uncertainty, as multi-model ensembles do.

A further possibility to quantifying the uncertainty in model uncertainty is by using stochastic parametrization. This recognises the mismatch between grid-scale resolved variables and sub-grid-scale unresolved processes and introduces random perturbations to sub-grid-scale tendencies. These have been shown to improve both the mean simulation characteristics of models and short-term weather forecast skill and more sophisticated techniques are currently under development (Palmer and Williams, 2008).

Uncertainty in climate-driven disease models

When a disease model is driven with predictions from a climate model, a fundamental source of uncertainty arises from the climate forcing. If the mapping of climate to disease outcome is known with exact precision, there would still be a range of outcomes in prediction due to the propagated uncertainty from the climate model. Conversely, if climate could be predicted with certainty uncertainty would remain in disease outcomes because of the uncertainty in the relation between climate forcing and disease.

The reality is that there is uncertainty in both halves of the modelling process. In climate modelling, as seen above, sources of uncertainties have been well described and quantified, however in general the uncertainty in climate-driven disease models, whilst having been acknowledged verbally, is not regularly quantified.

A large source of uncertainty in predicting disease output comes from the forcing from other systems such as landscape features, socio-demographic features and others. These can be built in more to models, however greater complexity can increase uncertainty. Predictions of each forcing necessarily come with their own uncertainties, compounding the result such that a range may increase far beyond the useful limit. Instead more useful predictions may be teased out by attempting to predict the effect of one system on another, holding everything else constant (or perhaps using a scenario based approach for different futures).

Furthermore, disease models are often parametrised, with parameters estimated from epidemiological lab or field work known to various degrees of certainty. For this reason it is possible to carry out sensitivity experiments in a way similar to perturbed physics climate ensembles. There is also structural uncertainty inherent in a disease model. Output from multiple-models for the same disease may then be collated to attain a better estimate of the true uncertainty in our knowledge of the link between climate and disease. However this is not widely practised, with output from single models generally used. This situation is now changing, with initiatives such as the Inter-Sectoral Impact Model Intercomparison Project (ISI-MIP, 2013), attempting to synthesise climate impact research from multiple models, with the goal of providing quantatitive estimates of climate impacts and their uncertainties. This project has recently begun to incorporate climate-driven disease models, particularly for malaria.

2.2.2 Defining an uncertainty framework

Various calls for and attempts to create a generalised framework for talking about uncertainty have been made (e.g. Janssen et al., 2005; Knol et al., 2009; Stainforth et al., 2007; Walker et al., 2003). These can potentially be useful for communication between modelling disciplines, since aspects of model uncertainty are known by different names to different users. However they are useful only when they are shared, otherwise the only result is an expansion of the number of names.

One of the most extensive typologies has come from Walker et al. and is discussed here (Walker et al., 2003). Their stated aim is to focus generally 'on the point of view of those providing information to support policy decisions, synthesising a wide variety of contributions on uncertainty in model-based decision support in order to provide an interdisciplinary theoretical framework for systematic uncertainty analysis'.

They adopt the general definition of uncertainty of being 'any deviation from the unachievable ideal of completely deterministic knowledge of the relevant system' or 'any departure from the ideal of complete determinism. They have attempted to harmonise existing typographies and they identify three dimensions of uncertainty; location, level and nature.

The location of uncertainty is where it exists within the model system and can then be subdivided into the five following categories;

- Context: an identification of the boundaries of the model, and thus the portions of the real world that are included and those which are not.
- Model uncertainty: associated with both the conceptual model (i.e. variables and their relationships that are chosen to describe the system) and the computer model. Can therefore be divided into two parts;
  - Model structure uncertainty, uncertainty about the conceptual model
  - Model technical uncertainty, uncertainty arising from the computer implementation of the model
- Inputs to the model: associated with forces that are driving changes in the system. Inputs can be divided into controllable and uncontrollable, depending on whether the decision maker has the capability to influence the values of the specific input variables.
- Parameter uncertainty: associated with the data and methods used to calibrate the model parameters
- Model outcome uncertainty: the accumulated uncertainty associated with the model outcomes of interest to the decision maker

Model outcome uncertainty is also called prediction error, as it is the discrepancy between the true value of an outcome and the models predicted value. If true values are known a formal validation exercise can be carried out to compare the true and predicted values in order to establish the prediction error as is carried out in this thesis.

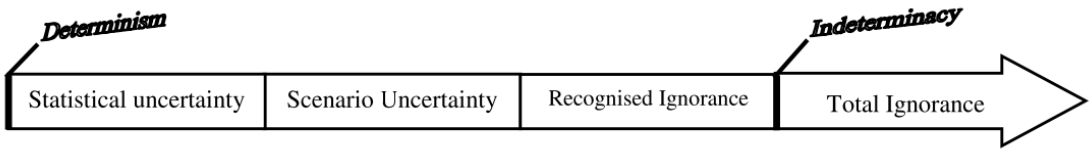


Figure 2.4: Levels of uncertainty (from Walker et al., 2003).

Level of uncertainty is where it manifests itself along the spectrum between deterministic knowledge and total ignorance. Level is a continuum, seen in figure 2.4. It ranges from determinism through to complete ignorance, passing through stages of statistical uncertainty, scenario uncertainty and recognised ignorance. Statistical uncertainty is where some quantification has been made. Scenario uncertainty is discussed in section 2.2.1 and it differs from statistical uncertainty, where the functional relationships are well described and a statistical expression of the uncertainty present can be formulated. Instead, scenario uncertainty implies that there is a range of possible outcomes, but the mechanisms leading to these outcomes are not well understood and it is not possible to formulate the probability of any one particular outcome occurring. There is a demarcation in the transition from statistical to scenario uncertainty, where a change occurs from a continuum of outcomes expressed probabilistically, to a range of discrete possibilities (Walker et al., 2003).

Recognised ignorance is when there is a fundamental uncertainty about mechanisms and functional relationships being studied. This occurs when neither functional relationships nor statistical properties are known and the scientific basis for developing scenarios is weak. Uncertainty due to ignorance can be further divided into reducible ignorance and irreducible ignorance. Reducible ignorance is that which may be resolved by conducting further research; irreducible ignorance is that which applies when more research cannot provide sufficient knowledge about significant relationships. Irreducible ignorance is also known as indeterminacy (Walker et al., 2003). Total ignorance is the other end of the scale from determinism and implies a deep level of ignorance.

Finally, the third dimension of uncertainty, nature, defines whether the uncertainty is due to the inherent variability of the phenomena being described or due to imperfection of our knowledge. This distinction is common in the uncertainty literature; common names given to the two states are *aleatory* and *epistemic* uncertainty (alternative names used in the literature are given in table 2.1).

Aleatory uncertainty	Epistemic uncertainty
Natural variability	Knowledge uncertainty
Random or stochastic variation	Functional uncertainty
Objective uncertainty	Subjective uncertainty
External uncertainty	Internal uncertainty
Statistical probability	Inductive probability

Table 2.1: Synonyms of aleatory and epistemic uncertainty, adapted from Baecher and Christian, 2000.

Aleatory uncertainty is associated with the natural variability in a process, the roll of a dice, that which can not be reduced by more knowledge. Epistemic uncertainty on the other hand is due to limited data and knowledge; with more work potentially this kind of uncertainty can be reduced. With perfect knowledge then, our uncertainty is identical to aleatory uncertainty. In reality however our knowledge is rarely perfect.

A general typology of uncertainty has been described; defining dimensions of location, level and nature. This is potentially useful for talking about computer model uncertainty in a general sense. Common language facilitates work between disciplines, though this language is not necessarily appropriate for communicating results to end-users who are often not specialists. It is important that effective communication of uncertainty is considered: this is the topic of the final section of this chapter.

2.2.3 Communicating quantified uncertainty

If uncertainty has been properly dealt with and quantified, yet is not communicated well, then the job is not complete. Without fully understanding forecast uncertainty a decision-maker can be dangerously overconfident. On the other hand if the comprehension level of the target of communication is not considered properly they may be overwhelmed by talk of uncertainty and paralysed against taking action.

A useful report regarding uncertainty communication has been prepared by Kloprogge et al., 2007, relating to issues and good practice in environmental management. Central within their discussion is the idea of 'progressive disclosure of information', which is defined as entailing:

> *the implementation of several layers of information, to be progressively disclosed; from non-technical information through more specialised information, according to the needs of the user* (Pereira and Quintana, 2002).

That is, uncertainty information is best offered gradually without overwhelming the recipient with all possible information.

The report also includes a comparison of three main methods of expressing uncertainty: verbal, numerical and graphical. Verbal has advantages such as the fact that most readers are more used to seeing or hearing uncertainty information communicated using words rather than numbers, and that words are better adapted to the level of understanding of lay audiences. This is also where its disadvantages lie; nuances may get lost and information oversimplified. The use of qualitative expressions rather than quantitative will also lead to different interpretations by different people in different settings.

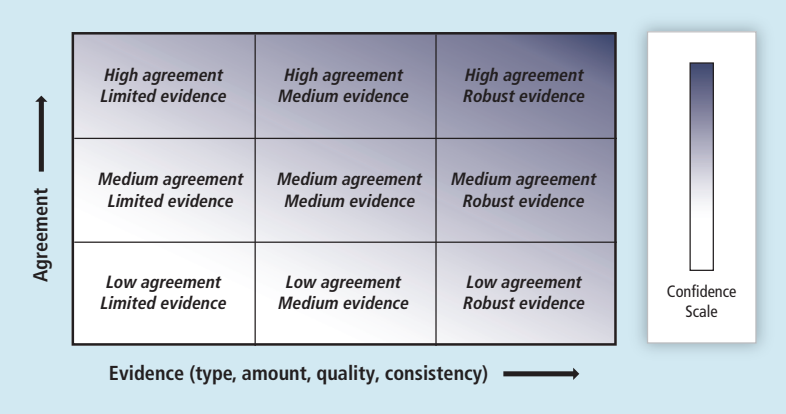


Figure 2.5: The relationship between uncertainty associated with a conclusion, and the agreement and nature of the evidence supporting it (from Mastrandrea et al., 2010).

The importance of precision with regards to language is well known by the IPCC. To encourage consistency, words relating to uncertainty were defined precisely at the outset of the AR4 report (IPCC, 2007), where they use the terms *high confidence*, *medium confidence* and *low confidence* to refer to 80%, 50% and 20% chance respectively. More recently, the instructions for authors of the upcoming AR5 report suggests a verbal-based definition of confidence along two dimensions: evidence and agreement (Mastrandrea et al., 2010, figure 2.5). Evidence relates to the type, amount, quality and consistency of evidence relevant to a conclusion, and is either and *limited*, *medium* or *robust*. Agreement can be said to either be *low*, *medium* or *high*. There is unavoidably some subjectivity to these definitions, however it enables a structured way of thinking about evidence and confidence: when there are multiple consistent independent lines of high-quality evidence (high agreement, robust evidence), confidence is highest, whilst it is lowest when evidence is sparse, low quality and inconsistent.

Numerical expressions of uncertainty have an advantage over words in that they are more specific, provided that readers understand how to interpret them and it is

possible to quantify uncertainty in the first place. If information is only presented with numbers, some readers will translate information into verbal expressions, potentially leading to miscommunication. Another thing to consider when presenting information numerically is that presenting numbers with more significant figures than is reasonable given the certainty of the result, suggests precision and confidence where there is none.

A final way to communicate uncertainty is to use graphics. They allow a lot of information to be summarised in a visually appealing form and are useful for non-specialists. If the method of representation of information is unfamiliar to a reader graphics have the disadvantage that time is necessary to understand the method of representation and to retrieve the uncertainty information. However when accompanied by verbal and numerical explanations graphics can be a very powerful method of communication, particularly so if the graphical concept is already familiar to the reader. This is the idea underlying weather roulette (Hagedorn and Smith, 2009), a conceptual framework for comparing forecast systems based on the well-known gambling game, making climate change projections comprehensible to a non-expert audience. Another visual representation of forecast information has been developed for seasonal forecasts, using a geometrical interpretation of probabilistic forecasts as a coloured triangle (Jupp et al., 2012). In summary graphics can a powerful form of uncertainty information: they capture the attention of the user, their non-verbal nature transcends language barriers and a well-designed graphic using colour well can conceptually represent uncertainty more fundamentally than words and numbers alone.

This concludes the literature review for the quantification of uncertainty in climate-driven disease risk. As mentioned before, validation of the driving models is a fundamental part of quantifying model-driven model forecast uncertainty. The following chapter briefly summarises technical information relevant to model validation, including descriptions of the observational datasets used in this project and the metrics used to measure model skill.

CHAPTER 3

Model validation

This chapter details the methodology employed to validate climate and climate-driven disease models. It is split into two sections; the first details the observational datasets used in this thesis. The second deals with the methods of validation of the model hindcasts, starting with a discussion of the levels of validation necessary when validating a climate-driven impact model. A mathematical description of metrics used here to measure model skill concludes the section. Details relating to model structure are not included here, instead these are left to the appropriate chapters.

3.1 Observations used in this study

When a sufficient number of operational forecasts to assess the performance of a forecasting model are not available[1], hindcasts can be produced. These are created by initializing models with past observations of the state of the climate system and integrating them forward in time. The resultant output can then be compared against observations over the corresponding period; this gives an estimation of confidence in the model. Validation is only possible when there is observational data to validate against.

It would be preferable to validate climate model hindcasts against observed meteorological parameters since there is a generally a high certainty that these are a good representation of reality. However this kind of data is not homogeneous in space and time, and has only become spatially dense enough to validate all of the grid points of a model in the past few decades, when satellite observations have enhanced coverage, since the 1970s. For this reason it is sometimes necessary to use reanalysis.

[1] This is normally the case if a model is new or has been developed for research purposes.

Reference dataset	Dates	Type	Reference
NCEP	1948-2010	Reanalysis	Kalnay et al., 1996
ERA40	1950-2002	Reanalysis	Uppala et al., 2005
ERA-Interim	1979-2010	Reanalysis	Dee et al., 2011
GPCP	1979-2010	Merged precipitation	Adler et al., 2003
20th Century	1871-2011	Reanalysis	Compo et al., 2011

Table 3.1: Summary of the reference datasets used in this study

Reanalysis products are created by assimilating observations from multiple sources (e.g. radiosondes, satellite products, surface pressure measurements etc.) into a climate model. This allows estimation of the complete past state of the atmosphere or ocean. By using proximate observations and knowledge of the mechanics of the climate system it is possible estimate the state at locations and times for when there were no observations. Creating quality reanalysis products is a demanding complex task, the reader is referred to the papers in table 3.1 for further details.

The advantage of using reanalysis is that the extent of spatial and temporal coverage can easily allow complete validation of hindcasts. The disadvantage is that there is lower confidence that reanalysis products represent reality, compared to observational data. If a model validates well against a reanalysis dataset it does not necessarily mean that it validates well against reality.

Accuracy of reanalysis data is considered to be higher in or near regions in space and time where more observations are assimilated, that is, generally in the recent past, and in Europe and the United States. By contrast there is a larger uncertainty associated with reanalysis when less observations exist (e.g. for Africa in the 1990s when civil unrest disrupted observing networks). Temporal and spatial scales are also important; reanalysis is less accurate at the local scale compared to the regional and synoptic scales. For example the amount of rainfall in one grid box occurring in one day in reanalysis has the lowest confidence whilst the total rainfall occurring over a wide region during a whole season is likely to be closer to reality.

Several reanalysis products have been used in this thesis; these are summarised in table 3.1 and described below.

- The **NCEP**/NCAR V2 reanalysis is issued by the National Centers for Environmental Prediction (NCEP) and the National Center for Atmospheric Research (NCAR) and covers the period 1948 to present (Kalnay et al., 1996). The model used is the NCEP global forecasting model, at roughly 250km x 250km horizontal resolution. Observations incorporated include radiosonde data, surface marine data from ships and buoys, and data from aircraft, surface

measurements and satellites. Various variables are available, including: surface temperature, surface pressure, latent flux at the surface and top-of-atmosphere fluxes, 10m wind and precipitation (total and convective). Data is available at sub-daily, daily and monthly resolution. In this study the fields for monthly average temperature and precipitation are used for validation for both the decadal models in chapter 4, and the seasonal models in chapters 5 and 6.

- The **ERA-40** reanalysis is produced by the European Centre for Medium-Range Weather Forecasts (ECMWF), covering 1958 to 2001 (Uppala et al., 2005). It has a 125km x 125km horizontal resolution, and is available at sub-daily, daily and monthly resolution. The model used is version Cy23r4 of the IFS forecast model used for operational weather forecasting, developed jointly by the ECMWF and Météo-France. Observations incorporated include data from radiosondes, aircraft data, surface synoptic observations, satellite data and radiation measurements from radiometers. Similar variables are available as in NCEP. Here, monthly fields of average temperature and precipitation are used for validation of the decadal models in chapter 4.

- **ERA-Interim** is the latest global atmospheric reanalysis produced by the ECMWF (Dee et al., 2011). It is an 'interim' reanalysis of the period 1979-present in preparation for the next generation of ERA-40. The IFS model is also used, though the version is Cy31r2, which was in use roughly five years after the ERA-40 version and so includes a large number of updates to the model. The spatial resolution is higher than ERA-40, roughly 80km x 80km, and the types of observations used are similar to ERA-40, as are the variables available from the model. In this project ERA-Interim has been used for tier-2 validation of the System 4-driven Liverpool Malaria Model (LMM) in chapter 7.

- The Global Precipitation Climatology Project dataset, **GPCP** (Adler et al., 2003), is used to validate precipitation for the decadal and seasonal models in chapters 4 to 6. GPCP is a monthly global analysis of surface precipitation, at roughly 250km x 250km resolution, covering 1979 to the present. It incorporates merged precipitation estimates from multiple types of satellite data, and surface rain gauge observations. The merging approach utilizes the higher accuracy of the low-orbit microwave observations to calibrate the more frequent geosynchronous infra-red observations and the combined satellite-based product is adjusted by the rain gauge analysis. Monthly average precipitation from GPCP is used for validation of both the decadal models in chapter 4 and the seasonal models in chapters 5 and 6.

- The **20th Century reanalysis** (Compo et al., 2011) spans 1871-2010 and assimilates

only surface pressure reports, using observed monthly sea-surface temperature and sea-ice distributions as boundary conditions. Spatial resolution is 2.2° x 2.2° and it is available at a six hourly timestep. This reanalysis was chosen because of its long time period, necessary for an exploration of the LMM in chapter 8; further details relating to the decision to use this dataset can be found in the methodology of that chapter.

3.2 Validation

The framework of validation followed here is the three-tiered framework described by Morse et al., 2005; a conceptual map is shown in figure 3.1. The first level of validation, tier-1, is the assessment of the quality of the climate model hindcasts with reference to either observations or gridded reanalysis.

When climate hindcasts are used to drive a climate impact model, such as the one for malaria used in this thesis, it is also necessary to validate the resulting output. This level of validation is referred to as tier-3 validation. This involves comparing predictions of disease prevalence from a climate-driven disease model with recorded epidemiological data. The collection method for epidemiological data should also be carefully considered; records from a large hospital in a region may be more reflective of the large scale prevalence of a disease compared to records from one local clinic, or records of incidence may simply represent the numbers of people presenting certain symptoms without a confirmed diagnosis.

Tier-3 validation is often not possible for a disease model, due to data constraints on observations. It may be that long time series of disease data simply do not exist (particularly the case for Africa), or data do exist but have quality issues. These issues could be related to discontinuities from changes in observation methods, or from the introduction of disease control programs. When tier-3 data of sufficient quality is not available then, tier-2 validation is possible. This involves comparing the climate model-driven model with the same impact model driven by climate observations. For tier-2 validation to be meaningful, it is required that the impact model be representative of the processes for that application field. Tier-2 does not validate the impact model *per se*; tier-3 validation is the only full test of a climate-driven impact model's skill.

For any of the three validation levels shown in figure 3.1, metrics can be used to measure skill. These metrics, or skill scores, are an important means of comparing the performance of forecasts to some baseline. Normally the mean, i.e. climatology, of a historical period is used as a reference. A wide range of skill scores are available in the

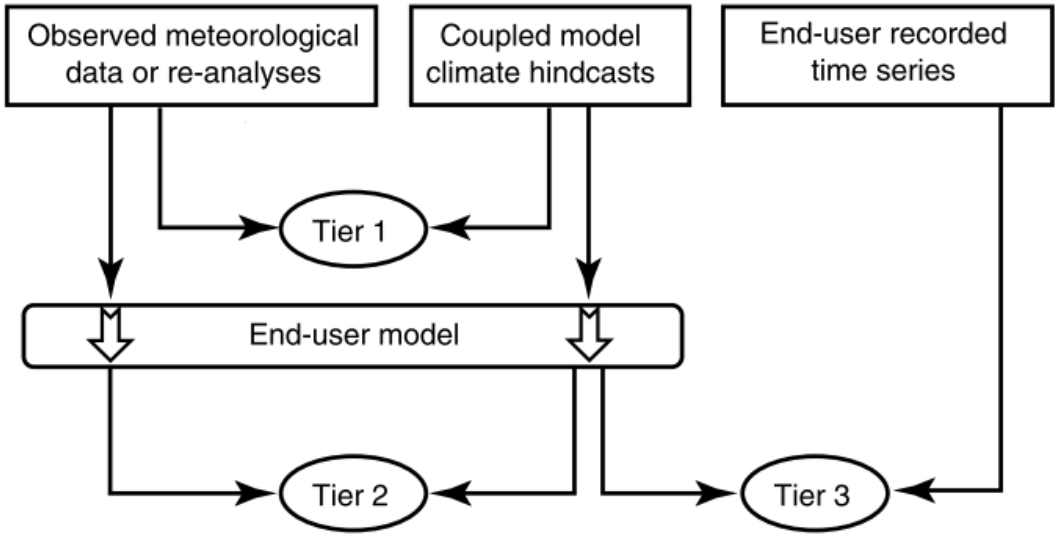


Figure 3.1: Schematic representation of the three-tier validation system (from Morse et al., 2005). Rectangular boxes represent sources of data, whilst ovals indicate the different types of validation.

literature, and a comprehensive review is given elsewhere (Jolliffe and Stephenson, 2003). A description of the skill scores used in this thesis follows.

All validation and visualisation was carried out using the NCAR Command Language, an interpreted language developed at NCAR (NCAR, 2013).

3.2.1 Pearson's product-moment correlation

The first score considered is the Pearson's product-moment correlation coefficient, r. It is a measure of the linear dependence between two sets X and Y, and can take any value ranging between -1 and 1. It is given by:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}, \tag{3.1}$$

where set X has n members X_i, with mean $\overline{X}$.

Significance levels are calculated at the 99% level, based on a two-tailed Student's t-test. Autocorrelation in the sample affects the significance levels (Wilks, 2011), and was taken into account by adjusting the effective sample size according to:

$$n' \cong n\frac{1-\rho_1}{1+\rho_1}\,, \tag{3.2}$$

where n and n' are the original and effective sample sizes respectively and ρ_1 is the lag-1 autocorrelation coefficient.

3.2.2 The relative operating characteristic area under curve

When assessing the performance of binary events and forecasts, a contingency table can be constructed which shows the combinations of events and forecasts. An example is given in table 3.2. The n forecast-observation pairs can be divided into the number of events which were correctly forecast (a), nonevents for which a false alarm was issued (b), events which were not forecast (c) and non-events for which the forecasting system correctly forecast a non-event (d).

Event forecast	Event observed	
	Yes	**No**
Yes	*Hit (a)*	*False alarm (b)*
No	*Miss (c)*	*Correct rejection (d)*

Table 3.2: Contingency table for binary events and forecasts

From this table, numerous skill scores can be calculated, as described by Jolliffe and Stephenson, 2003. For example, the hit rate, H, is given by

$$H = \frac{a}{a+c}\,, \tag{3.3}$$

that is, the proportion of occurrences of the event that were correctly forecast. It is a measure of the ability of the forecasting system to correctly forecast events; a perfect forecast has $H = 1$. However, this score alone is not a reliable measure of forecast skill; it does not measure false alarms, so apparent performance be improved by simply issuing more 'yes' forecasts. For this reason H is usually used along with the false alarm rate, F, given by,

$$F = \frac{b}{b+d}\,. \tag{3.4}$$

False alarm rate is the complimentary score, giving the proportion of non-occurrences which were incorrectly forecasted. A perfect forecast has $F = 0$ and again, F cannot be used alone since it can be improved by simply increasing the number of rejections issued.

In this project, an 'event' associated with a variable (i.e. temperature, precipitation or malaria incidence) is defined as when a it falls either below or above the upper tercile of climatology. This is used to indicate high and low anomalies.

A method to assess the ability of a forecast system to correctly detect events arises from signal detection theory and is known as the relative operating characteristic (ROC) curve. The method assumes that an 'event' is preceded by some signal in the data superimposed on a background of noise, whilst a non-event is preceded by noise alone. The forecasting system (or forecaster) issues a 'yes' forecast if the weight of evidence is sufficiently great. This is illustrated in figure 3.2.

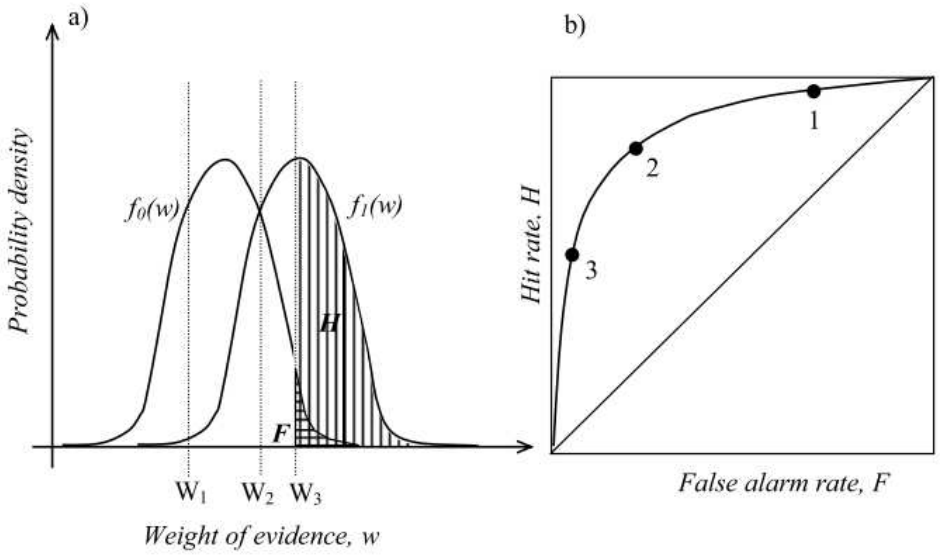


Figure 3.2: Schematic of (a) basic signal detection model and (b) corresponding ROC curve, from Jones, 2007, based on figures from Jolliffe and Stephenson, 2003. $f_0(w)$ and $f_1(w)$ are the distributions of the "weight of evidence" w before a non-event and an event respectively. The hit rate H and false alarm rate F are the shaded areas in (a). The ROC curve in (b) then plots hit rate against false alarm rate and is created by varying the threshold. Points 1-3 in (b) correspond approximately to the thresholds W_1 W_2 and W_3 in (a).

The 'weight of evidence' variable, w, has probability distribution $f_0(w)$ before a non-event and $f_1(w)$ before an event. The better the forecasting system, the further apart the two distributions. A 'yes' forecast is issued if the weight of evidence exceeds some decision threshold W. The hit rate and false alarm rate correspond to the shaded regions in figure 3.2a, using a threshold of W_3.

The ROC curve in figure 3.2b is a plot of hit rate against false alarm rate, and can then be drawn as a graph of H against F, with points created by varying W. For low W, H and F are high, and then decrease as W increases. For a perfect forecasting system (no overlap

between the two distributions), the ROC curve describes the square, moving vertically from (0,0) to (0,1) and then horizontally to (1,1). For an unskilful forecasting system (full overlap between the two distributions) H and F change simultaneously and the ROC curve lies along the diagonal $H = F$.

A measure of skill commonly used in meteorology and employed here for validation is the area under the ROC curve (ROC AUC) which is equal to the probability that given an event and a non-event, the forecasting system will correctly categorise the two (Jolliffe and Stephenson, 2003). A perfect forecast has a ROC AUC of 1 whilst a forecasts showing no skill over climatology have an area of 0.5.

Significance levels for the ROC AUC are calculated by a comparison to the Mann-Whitney U-statistic (Mason and Graham, 2002). The Mann Whitney test is non-parametric, and tests to see which one of two independent samples tends to have larger values than another. The statistic follows a known probability distribution, which can be used to define the statistical significance of the ROC area.

The method employed here follows Mason and Graham, 2002, where the sample of forecast probabilities of all the events is compared with the sample of all forecast probabilities of non-events. The number of events, n_1, and non-events, n_2, are then used to calculate critical values of a two-tailed Mann-Whitney U. A two-tailed test is used because random data gives on average a ROC AUC of 0.5, and a forecast can either be perfect with a ROC AUC of 1, or perfectly wrong with a score of 0.

Once a value of U is found, this can be related to the ROC AUC (here R) by the following relationship:

$$R = \frac{\mathrm{U}}{n_1 n_2} \,. \tag{3.5}$$

As an example, looking at upper tercile events in a sample of 30 forecasts gives $n_1 = 10$ and $n_2 = 20$. At 95% significance the critical U value for these sample sizes is 55, giving a ROC AUC as 0.275. This then suggests that ROC AUC below 0.275 or above $(0.5 + 0.275) = 0.775$ is significant at the 95% level. Note that this is rough approximation to significance; when there are instances where forecast probabilities take the same value in the sample a correction must be taken into account, but when the number of ties is small this correction is not large (Mason and Graham, 2002).

For reference, 95% and 99% significance levels based on this method are shown in table 3.3 for ROC AUC for upper tercile events for different sample sizes.

Sample size	95% significance	99% significance
15	0.83	0.93
20	0.78	0.87
35	0.71	0.77
50	0.67	0.73
100	0.62	0.66

Table 3.3: Significance levels of ROC AUC for tercile events for various sample sizes.

3.2.3 Reliability diagrams

The ROC AUC measures the ability of a forecast system in detect an 'event'. Reliability diagrams provide additional information; they measure how closely the forecast probabilities of an event correspond to the actual chance of observing it. An example of a reliability diagram is given in figure 3.3.

The reliability diagram groups forecasts of all gridpoints in a region into bins according to their forecasted probability (horizontal axis). For each subset of forecasts separately, the frequency with which the event is observed to occur is plotted against the vertical axis. For perfect reliability the forecast probability and the frequency of occurrence should be equal, and the plotted points should lie on the diagonal (solid line in the figure). Thus, for all of the occasions when a perfectly reliable forecast system states an event will occur with a probability of 80%, the event will indeed occur on 80% of those occasions.

In the figure 3.3, the reliability curve has a positive slope, indicating that as the forecast probability of the event occurring increases, so too does the chance of observing the event. The forecasts therefore have some reliability. However, the slope is greater than one, indicating less than perfect reliability. The information contained in reliability diagrams may therefore be used to make approximate corrections to the raw model forecast probabilities.

Also shown on the diagram is a histogram of distribution of forecasts. This is also known as a sharpness diagram and shows the relative frequency with which the event has been predicted (over the reference period and at all gridpoints) with different levels of probability. A forecast system which is capable of predicting events with probabilities different from the observed event frequency is said to have 'sharpness'. A forecast system with no sharpness exhibits a frequency peak near the climatological frequency, indicating that the majority of forecasts predict the event with a probability near the climatological frequency. For planning purposes such forecast systems offer little value over climatology. In the example of figure 3.3, the sharpness diagram shows that the largest category is a 0% forecast, with forecast probabilities distributed roughly

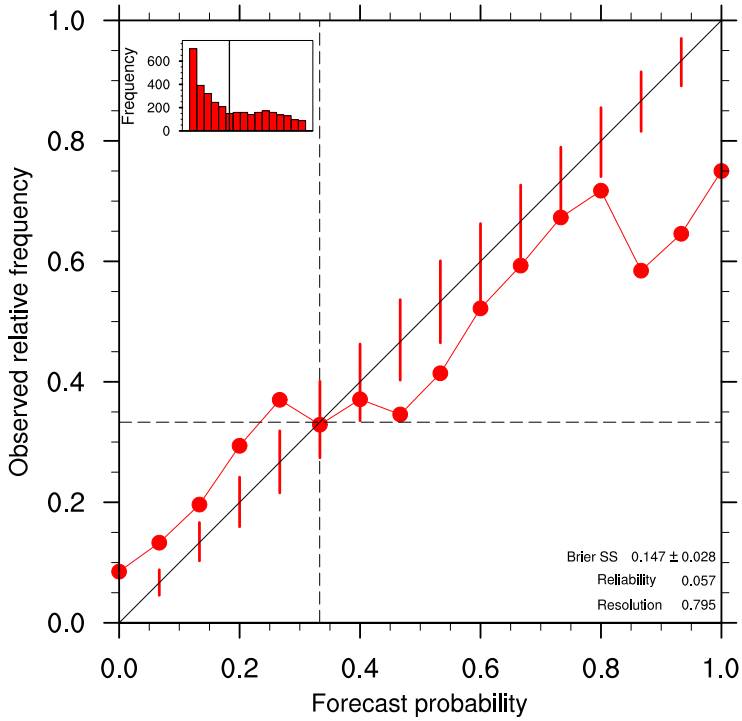


Figure 3.3: An example of a reliability diagram for an upper tercile event. Dashed lines indicate the climatological frequency.

evenly across the rest of the bins. Thus this particular system has some sharpness; probabilities are not clustered around the climatological frequency (here 33% as it comes from prediction of an upper tercile event).

Finally the significance of the reliability diagram is estimated by employing consistency bars (Bröcker and Smith, 2007). These give an idea of how likely a observed relative frequency is under the assumption that the predicted probabilities are reliable. To calculate bars a 1,000 member resampling was performed and the 5-95 percentiles of the distribution calculated for each bin point (following the method described in Bröcker and Smith, 2007). This percentile range is then plotted as error bars for each point on the reliability curve. A point inside the consistency bars does not indicate categorically if a forecast is reliable or unreliable, but it is unlikely (<5% chance) that a point outside the bars comes from a perfectly reliable forecast system. Points inside narrow consistency bars can be said with more confidence to come from reliable systems than those inside large consistency bars (with small populations).

The Brier score and its decomposition is also included on these reliability plots. A description of this score follows.

3.2.4 Brier skill score

The Brier score is a measure of the mean square probability error for binary events and is given by

$$B = \frac{1}{n}\sum(p_i - x_j)^2 \,, \tag{3.6}$$

where p_i is the forecast probability of forecast i and x_j is equal to 0 if the event did not occur and 1 if it did (Brier, 1950). The Brier skill score (BSS) is the Brier score calculated relative to a reference forecast, usually a constant forecast of the climatological probability of the event, given by

$$BSS = 1 - \frac{B}{B_{ref}} \,. \tag{3.7}$$

A perfect forecasting system has $BSS = 1$ ($B = 0$), whilst a forecast system equal in skill to the reference has $BSS = 0$. Forecasts with skill lower than climatology have a negative BSS. Analytical error bounds on the BSS are been calculated following the method in Bradley et al., 2008; the reader is referred to the paper for the (mathematically involved) method.

The Brier score can be decomposed into resolution and reliability components (Jolliffe and Stephenson, 2003). These are given by:

$$B_{rel} = \frac{E_q[(q - f(q))^2]}{s(1-s)} \tag{3.8}$$

and

$$B_{res} = 1 - \frac{E_q[(f(q) - s)^2]}{s(1-s)} \,, \tag{3.9}$$

where $f(q) = p(X = 1|q)$, that is, the conditional probability that the event occurs given the forecast probability q, and s is the climatological base rate of the event (i.e. the long-term average frequency). These components are both positively orientated, with a zero value indicating perfect reliability or resolution, and large values indicating poor scores. The BSS can be expressed in terms of resolution and reliability scores,

$$BSS = 1 - B_{rel} - B_{res} \,. \tag{3.10}$$

Reliability and resolution are two of the most important attributes of a forecast system

(Jolliffe and Stephenson, 2003). Reliability is described in the previous section and is considered as the lesser important of the two attributes as it can be improved by simply calibrating the forecast probabilities issued against observed frequencies using a reliability diagram. However, resolution cannot be improved by calibration and is considered an intrinsic measure of the value of a forecasting system. It measures how much conditional probabilities differ from the climatic average; a system with poor resolution will always forecast an event with the climatological probability, whilst a system with good resolution will issue a wider range of probabilities and can potentially better *a priori* identify situations that lead to the occurrence or non-occurrence of an event (Jolliffe and Stephenson, 2003).

3.2.5 Value

Turning from general forecasting diagnostics toward metrics more useful to decision makers, here the potential economic value of forecasts is described. This measure is also sometimes called the relative value or just the value. It has been used by numerous authors for assessment of meteorological forecasts, as well as being applied in impact studies such as hydrology and health (Morse et al., 2005; Palmer, 2000; Richardson, 2000; Zhu et al., 2002).

The principal behind economic value is the cost/loss model (Murphy, 1969), a simple decision model which assumes that decision makers can respond to a forecast event by taking action at a cost C in order to avoid a loss L. Potential expense outcomes are shown in the contingency table in figure 3.4.

Action taken	Event occurs	
	Yes	**No**
Yes	C	C
No	L	0

Table 3.4: Cost/loss model showing the potential expense if a decision maker chooses to act in anticipation of an uncertain event.

A decision maker might choose to always take action (with cost C) to avoid a loss, with average expense

$$E_{always} = C \,, \tag{3.11}$$

or they can choose never to act, incurring an average expense of

$$E_{never} = sL\,, \tag{3.12}$$

where s is the probability of the event occurring (in this case of tercile category forecasts $s = 1/3$), and L is the loss. A rational strategy in the absence of information then is to either always or never act (depending on which of E_{always} and E_{never} is lower). This allows the definition of the average 'climate expense' as:

$$E_{climate} = min(c, sL)\,, \tag{3.13}$$

which gives a baseline for measuring the value of a forecast. It is the expected expense in the absence of forecast information and a forecast is useful if it reduces this mean expense. Value, V, of a forecast system is therefore defined as the reduction in mean expense provided by the forecasting system, relative to the reduction obtained if one had access to perfect forecasts:

$$V = \frac{(E_{climate} - E_{forecast})}{(E_{climate} - E_{perfect})}\,. \tag{3.14}$$

$E_{perfect}$ here is the incurred expense with access to perfect forecasts (i.e. with perfect knowledge of the future). In this case one would only take action every time an event occurs (each time with expense C) and would experience no losses. Thus;

$$E_{perfect} = sC\,. \tag{3.15}$$

The mean expense of a forecast system can be found by multiplying the cost/loss matrix with the system-specific contingency table (table 3.2) and expressing as a function of H, F and s. This gives:

$$E_{forecast} = sCH + sL(1 - H) + (1 - s)CF\,. \tag{3.16}$$

The first term here responds to the expense associated with correctly predicted events, the second with misses and the third with false alarms. There is no term for correct rejections, as there is no expense associated with this outcome. With some manipulation, equation 3.16 can be expressed as:

$$E_{forecast} = F(1 - s)\alpha - Hs(1 - \alpha) + s\,, \tag{3.17}$$

where F and H are the false alarm and hit rate respectively and $\alpha = C/L$ is the ratio of cost to loss. It is then possible to write an equation for V by substituting the preceding equations into equation 3.14:

$$V = \frac{min(\alpha,s) - F(1-s)\alpha + Hs(1-\alpha) - s}{(min(\alpha,s) - s\alpha)} . \tag{3.18}$$

By using this equation, curves of V vs. α can be plotted over the range $\alpha = [0,1]$. Where $V > 0$ the forecast system has value, and a value of $V = 1$ corresponds to a perfect forecast.

Value curves can be plotted for a range of decision thresholds (i.e. the forecast probability necessary to trigger a decision to act); an example is given in figure 3.4. After identifying their cost/loss ratio a user can use the set of value curves to select a decision threshold for a forecasting system which maximizes their value.

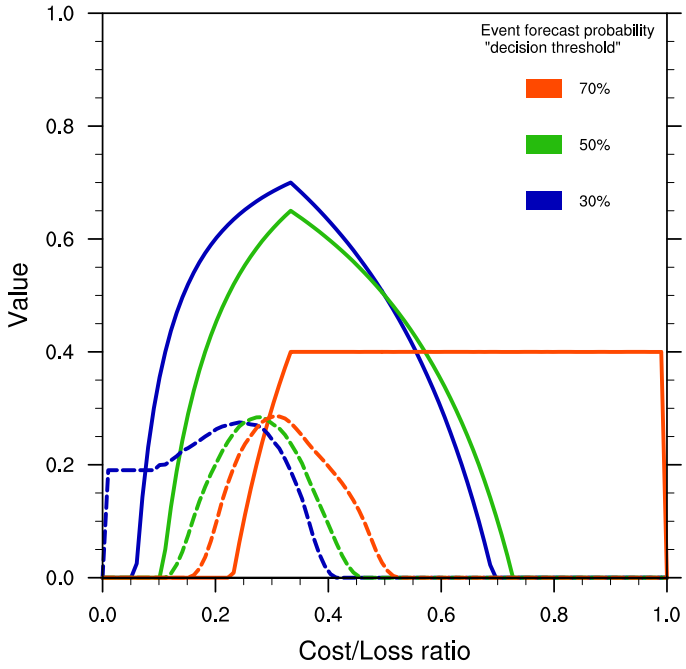


Figure 3.4: An example of value curves for different decision thresholds. Dashed lines indicate 95% significance levels for each decision threshold.

Note that it is sometimes the case that value is independent of cost/loss ratio (creating a horizontal line on the value curve): this occurs when either $\alpha < s$, the decision threshold is low and the hit rate is one, or for $\alpha > s$ when the threshold is high and the false alarm rate is zero. This can be shown by substituting either $min(\alpha,s) = \alpha$ and $H = 1$ or $min(\alpha,s) = s$ and $F = 0$ into equation 3.18 above.

No method for calculating significance levels for economic value curves was found in

the literature, so one was created. The method involved calculating an upper limit to the range of value curves one would expect if the forecasts came from a random process. To calculate this limit, two time series of pseudo forecast-observation pairs were created of the same length as the time series for which significance curves are required. 'Forecasts' consist of continuous random numbers in the set $[0,1]$, whilst the 'observations' are a random binary series of 0 (non-events) and 1 (events), where

$$\frac{N_1}{N_0} = s\,, \tag{3.19}$$

where N_1 is the number of 'events' and N_0 is the number of non-events. That is, the number of events in the sample is made to agree with the baseline frequency. For example when looking at tercile events over 30 years, 10 years are randomly selected as observed events. For each decision threshold separately these pseudo forecast-observation pairs are then used to calculate a value curve and this is repeated 10,000 times. The smoothed 95th percentile is then used as the 5% significance level.

The significance curve calculated depends on the length of the time series, on the average frequency of the event s (equal for upper and lower tercile events), and on the decision threshold. In the plots of value curves in the thesis, colours are used to distinguish value curves for different decision thresholds, with the calculated value curve shown with a solid line and the significance level indicated with a dashed line of the same colour.

3.2.6 Potential predictability

Potential predictability, PP, is used only in chapter 4. PP gives an idea about the magnitude of the predictable signal with respect to the total signal and it quantifies the fraction of the variance of a climate variable due to a common external forcing compared with the internal variability of the coupled atmosphere-ocean climate system. This is estimated in the model world and does not provide any information about the realism of the atmospheric signal. As such it is only useful alongside measures of skill such as correlations. A complete description of the method used to calculate potential predictability is given elsewhere (Storch and Zwiers, 1984), a more specific description follows here.

Let n represent the number of simulations in the ensemble ($j = 1, n$) and N the number of years for each member ($i = 1, N$). Let χ represent an atmospheric variable (rainfall or temperature). The value of this variable at a time step i for a member j is given by:

$$\chi_{ij} = \mu_i + \epsilon_{ij} \,, \tag{3.20}$$

where μ_i is the atmospheric response to the common external forcing and ϵ_{ij} the chaotic fluctuation related to each member. The assumption is made that the external variability does not interact with the internal variability and as a consequence the total variance of the signal equals the sum of these two components (external/internal):

$$\sigma^2_{TOT} = \sigma^2_{EXT} + \sigma^2_{INT} \,. \tag{3.21}$$

This method assumes that the internal variability is Gaussian. It is calculated as the average over the whole time period as the quadratic deviation of each member from the ensemble mean:

$$\sigma^2_{INT} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{n-1}\sum_{j=1}^{n}\left(\chi_{ij} - <\chi_i>\right)^2\right] \,, \tag{3.22}$$

with $<>$ representing the ensemble mean and

$$<\chi_i> = \frac{1}{n}\sum_{j=1}^{n}\chi_{ij} \tag{3.23}$$

representing the ensemble mean for a given year i. The variance of the ensemble mean, σ^2_{EM}, is given by:

$$\sigma^2_{EM} = \frac{1}{N-1}\sum_{i=1}^{N}(<\chi_i> - <\overline{\chi}>)^2 \,, \tag{3.24}$$

with $\overline{\chi}$ representing the time mean of χ and

$$<\overline{\chi}> = \frac{1}{nN}\sum_{i=1}^{N}\sum_{j=1}^{n}\chi_{ij} \tag{3.25}$$

the ensemble mean for the whole integration period. The variance due to the common external forcing is then given by:

$$\sigma^2_{EXT} = \sigma^2_{EM} - \frac{1}{n}\sigma^2_{INT} \,. \tag{3.26}$$

This equation shows that the ensemble mean is an estimation of the common external forcing; this estimation is more accurate when the number of simulations in the ensemble n is large. PP can then be defined as the ratio of the external variance to the total variance:

$$PP = \frac{\sigma^2_{EXT}}{\sigma^2_{TOT}} \,, \tag{3.27}$$

with PP expressed as a percentage.

This concludes the section on theory related to model validation. Reanalysis and observational datasets against which models are compared have been described, along with the metrics used to measure skill. Results follow, beginning with part one and the validation of decadal climate forecasts.

Part I

The skill of climate forecasts

CHAPTER 4

Validation of the ENSEMBLES stream 2 decadal climate hindcasts.

This chapter considers climate models run in decadal prediction mode, and the potential to use their output to forecast disease. The work is based around a paper published in Environmental Research Letters (MacLeod et al., 2012), validating the first multi-model decadal climate model hindcast set, produced as part of the ENSEMBLES project. Some elements of the paper are reproduced here exactly, with others enhanced to include work and description left out of the paper.

Analysis focuses on the skill in prediction of temperature and precipitation - variables important for impact prediction. Whilst previous work on this dataset has focused on the skill in prediction of multi-year averages (Oldenborgh et al., 2012), here the focus is on skill in prediction of annual and seasonal averages.

4.1 Introduction

The ability to forecast the evolution of climate on decadal timescales would prove a useful tool to aid climate change adaptation. Information on this timescale is important for adaptation as it is a key planning horizon for governmental and non-governmental organisations, businesses, and other societal entities (Cane, 2010).

Climate model projections are generally initialised from randomly selected pre-industrial states, so the variability in projections is not synchronised with observations (Meehl et al., 2009). However recent work has shown that it is theoretically possible to improve skill in predicting some aspects of global and regional climate over a decade in advance by initializing climate models with observations. This

has led to further research on decadal climate prediction, along with claims about its potential to anticipate climate impacts (Keenlyside and Ba, 2010; Meehl et al., 2009; Murphy et al., 2010). Despite this it has not yet been demonstrated that decadal predictions have sufficient predictive skill to be used operationally.

Decadal climate modelling has evolved from seasonal climate modelling, which is able to provide useful forecasts on regional scales for temperature and precipitation in certain locations, particularly the tropics, and is regularly used operationally. On seasonal timescales predictability largely arises from slowly varying sea surface temperatures and major modes of variability such as the El Niño Southern Oscillation (ENSO) in the coupled climate system (Palmer and Hagedorn, 2006b). On decadal scales, predictability is believed to arise at least partly from lower-frequency climate modes and forced boundary conditions. Climate modes which may potentially offer predictability on decadal timescales include the Atlantic Multidecadal Oscillation, a basin-wide fluctuation of sea surface temperatures in the North Atlantic with a periodicity of around 70 years (Schlesinger and Ramankutty, 1994) and the Pacific Decadal Oscillation, a similar pattern of climate variability in the Pacific (Mantua et al., 1997). Predictability from boundary conditions comes from anthropogenic (e.g. greenhouse gas, aerosols) and natural (e.g. volcanic, solar) sources (Keenlyside and Ba, 2010).

To explore the potential for skilful decadal prediction, the first multi-model decadal hindcast set was made as part of the ENSEMBLES project, covering the last 50 years (Van Der Linden and Mitchell, 2009). The models used are state-of-the-art coupled ocean-atmosphere global circulation models, integrated forward for ten years from ten start dates distributed throughout the hindcast period, 1960-2005. Previous work with this hindcast dataset has looked at skill in annual average temperature and precipitation for the first year from initialization and in four year average blocks thereafter (Oldenborgh et al., 2012), focusing on North Atlantic and Pacific sea surface temperatures and on global average temperature. However, climate impacts depend on time and space scales shorter than this; regional level sub-annual climate is the principal driver of climate impacts and interannual variations in seasonal temperature and precipitation over small regions can have significant socio-economic impacts, for example on agriculture, health and other sectors (Washington et al., 2006). The focus here then is the question: can the ENSEMBLES decadal hindcasts anticipate interannual variations in temperature and precipitation on the time and space scales relevant to climate impacts?

Annual and seasonal averages are considered at different regional scales to see if the predictions available from these decadal hindcasts can drive impact models with skill.

Details of models and validation methods are described in section 2, section 3 contains results and discussion and conclusions can be found in section 4.

4.2 Methodology

The ENSEMBLES multi-model decadal stream 2 hindcasts consist of four forecast systems: IFS33r1, HadGEM2, ARPEGE4.6 and ECHAM5, developed at ECMWF, UK Met Office, CERFACS and IFM-GEOMAR respectively. Details of the models with further references can be found elsewhere (Oldenborgh et al., 2012). Three members for each model were run for ten years starting on 1st November 1960, 1965 and every five years thereafter until 2005, giving nine hindcast time blocks and one which extends into the future (Van Der Linden and Mitchell, 2009). Throughout this analysis annual (November to October) and seasonal averages are considered, specifically boreal winter and summer (hereafter referred to as DJF and JJA respectively). To validate the hindcasts, multiple reference datasets have been used: NCEP and ERA40 for temperature, and NCEP, ERA40 and GPCP for precipitation. Details of these datasets can be found in chapter 3. Multiple references were used since reanalysis has uncertainty associated with it and validating with multiple references evaluates the robustness of results.

Each of the models has a temperature drift, dependent on the lead time from initialisation. To counter the confounding effect that this drift would have on the statistics, it was removed point-wise. To do this the lead-time dependent drift was first calculated by averaging all of the hindcasts from each of the four models separately to create four ten year time series. From these the average of reference datasets averaged over the same periods was subtracted (using NCEP when correlating against NCEP and ERA40 when correlating against ERA40). This created four drifts - one for each model, for every grid point. These lead-time dependent drifts were then subtracted from every member in the hindcasts corresponding to each model. Cross-validation was not used, and all subsequent analysis uses the drift-corrected data.

The similarity between hindcasts and observations at annual and seasonal scales was tested for by calculating Pearson's product-moment correlation coefficients between the ensemble mean and each reference dataset, for annual and for seasonal averages at all lead times. Two-tailed 99% significance levels were calculated based on Student's t-test, dependent on sample size which varies with reference dataset. Furthermore, serial correlations were taken into account by adjusting the effective sample size according to:

$$n' \cong n\frac{1-\rho_1}{1+\rho_1}\,, \tag{4.1}$$

where n and n' are the original and effective sample sizes respectively and ρ_1 is the lag-1 autocorrelation coefficient from observations (Wilks, 2011).

Global maps of temperature correlations have been plotted, before and after detrending, along with precipitation correlations (detrended precipitation correlations are not shown since the trend in precipitation over the hindcast period was not significant). To detrend, the linear regression averaged across all time blocks was subtracted point-wise from each individual hindcast member, doing the same for observations (i.e. separately for each block). Only correlations with NCEP (for temperature) and GPCP (for precipitation) are contained in this chapter, with figures corresponding to the other reference datasets presented in appendix A and commented in the text.

After assessing the skill of the ENSEMBLES decadal hindcasts in reproducing observed interannual temperature and precipitation there remains the question of the signal to noise ratio, namely to what extent predictable regional variations might rise above noise from uncertainties in the forced response of the simulated coupled climate system. Analysis of variance tests are generally employed to separate the total variability for a given climate variable into an unpredictable component (mainly arising from atmospheric dynamics and the ocean-atmosphere coupling at short time scales) and a potentially predictable component due to the slow varying external boundary forcing (anthropogenic such as greenhouse gases and natural such as volcanic). For both temperature and precipitation, a one way analysis of variance (Storch and Zwiers, 1984) was applied to the decadal ensemble hindcast to quantify the fraction of the variance due to the common external forcing compared to the internal variability of the coupled atmosphere-ocean climate system. This ratio is also known as potential predictability (PP). A complete description of the calculation of PP is contained in section 3.2.6.

For the decadal simulations, all time blocks were concatenated as a single time dimension before performing this analysis, to increase the sample size. That is, each of the ten ten-year forecasts were combined into one timeseries. Anomalies were then calculated with respect to each model ensemble mean before performing the analysis, to avoid deflating the PP due to different model biases. PP is expressed as a percentage and gives an idea about the magnitude of the predictable signal with respect to the total signal. This is estimated in the model world and does not provide any information about the realism of the atmospheric signal, which is why it is only useful alongside measures of model accuracy, such as correlations.

The ability of the models to simulate trends over different regions was also explored. Drift-corrected hindcasts were averaged spatially over 18 regions and subsequently trends were calculated by a linear regression in the hindcasts and in the observations at each start date. Trends were calculated for each ensemble member separately. This was repeated for multiple trend lengths, from five up to ten years, each starting from the first year (i.e. year 1-5, 1-6, 1-7 etc.) and was calculated for trends in both annual and seasonal averages. Subsequently, Spearman's rank correlations between all the trends simulated by one model and one reference dataset were calculated, along with significance levels at the 90%, 95% and 99% level. In order to display the maximum amount of information possible only the exceedance of significance levels are indicated for each of the models, rather than providing the exact value of the correlation.

4.3 Results

All decadal results are summarised in table 4.1. Subsequently results for each score are described in separate sections.

4.3.1 Biases

As a starting point the temperature biases for the four models and the multimodel ensemble averaged over all start dates and all lead times is shown in figure 4.1. For reference, the observed climatology for the hindcast period from the NCEP, ERA40 and ERA-Interim reanalysis datasets are presented in appendix A in figure A.1, as well as model climates in figure A.2. When looking at annual averages, there is relative agreement between the reanalysis datasets (figure A.1), whilst the same is true for DJF and JJA averages. Temperature simulation in the decadal models are therefore validated against NCEP, since it covers the entire hindcast period.

Each of the decadal models has a different bias, which varies in space and time. The largest biases in ECWMF are a $1-2^{\circ}C$ cold bias over the ocean, and a warm bias over northern hemisphere land, reaching maximum of around $3-4^{\circ}C$ over northern Canada and the Tibetan Plateau (figures 4.1a - 4.1c). For the rest of the globe, biases are smaller, with most of Africa showing a bias of under $\pm 1^{\circ}C$. These biases are roughly equal for annual, DJF and JJA averages. The UKMO model has its largest biases over eastern Canada for annual and DJF average (over $-5^{\circ}C$, figures 4.1d - 4.1f). This bias reduces in JJA, with a similarly large warm bias over the USA. Over the tropics the bias is warm, around $3^{\circ}C$. For UKMO the biases are less similar between seasons; DJF average bias is much colder whilst the JJA bias is much warmer, particularly in the northern hemisphere.

CERFACS has a large cold bias over the whole world (figures 4.1g - 4.1i). This is over $-5^{\circ}C$ for large areas, and is stronger over land than the ocean. The only area of warm bias is over eastern Canada in DJF, where the bias reaches over $5^{\circ}C$, and over northern Brazil at all temporal averages. IFM-GEOMAR has on average the lowest temperature bias of all the models, with a bias of under $1^{\circ}C$ over most regions of the globe, excluding Antarctica (figures 4.1j - 4.1l). Areas of larger bias include a warm bias over the Tibetan Plateau, a cold bias over northern Asia and a warm bias over North America in DJF. The temperature simulation in Antarctica (and to a lesser extent, over the Arctic) has large biases in all models - exceeding $\pm 5^{\circ}C$. An obvious reason for this is poor model simulation of the dynamics in these regions. However, reanalysis data is highly uncertain in this region due to the lack of temperature measurements - the target

Temperature

Model biases

- Warm bias over most of the Americas for all models, in other regions variability between models
- CERFACS has largest bias; more than $5°C$ too cold over large areas of the globe

Ensemble mean correlation

- Before detrending: significant correlation everywhere with maximum over the tropics, higher for annual than seasonal averages
- After detrending: no significant correlation anywhere

Potential predictability

- Higher for annual than seasonal averages
- Highest over the tropics, particularly the Maritime Continent. Low elsewhere

Trend correlations

- Significant correlation for all models for global trends at 5-7 years
- For smaller spatial averages than global some models show significance for multiple trend lengths; regions listed in the text (also see figure 4.7a)

Precipitation

Model biases

- All models overestimate annual precipitation over the Maritime Continent and JJA precipitation over West Africa
- Not enough rainfall over the Amazon and India in JJA in all models

Ensemble mean correlation

- Below significance everywhere, for annual and seasonal averages

Potential predictability

- Much lower than temperature
- Mostly lowest over land

Trend correlations

- No correlation above significance for any region, for annual or seasonal averages

Table 4.1: Summary of all results for decadal prediction

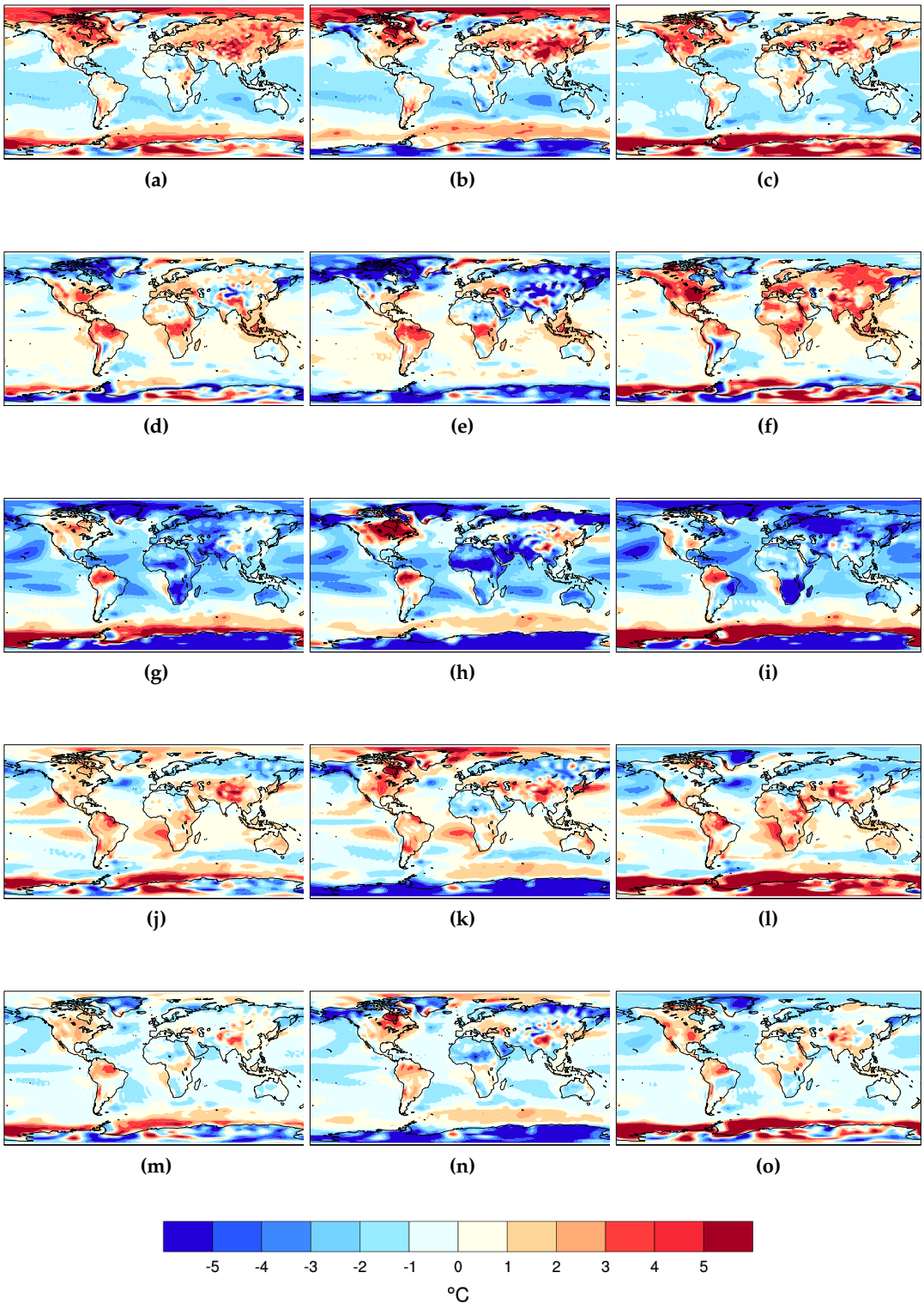


Figure 4.1: Annual, DJF and JJA (left, central, right column) temperature biases from ENSEMBLES stream 2 decadal models. For ECMWF (a-c), UKMO (d-f), CERFACS (g-i), IFM (j-l) and the multimodel mean (m-o). NCEP used as a reference. Bias is calculated by averaging over all start dates and lead times.

which model temperature is measured against does not necessarily reflect reality. In any case, the lack of civilization in these regions means this area is a low priority for climate-driven disease predictions, and accordingly predictions here are not further discussed.

Looking now at the bias of the multi-model ensemble (figures 4.1m - 4.1o), showing the average bias across the four models. Firstly the oceans are too cold at all timescales, with an average bias of around $1^{\circ}C$. Secondly the Americas have warm biases, for annual and seasonal average. Finally there is a shared cold bias across the north of Africa.

Precipitation biases are shown in figure 4.2, whilst observed and model climatologies can be found in the appendix, in figures A.3 and A.4. The general spatial pattern of precipitation is roughly consistent between the reference datasets (figure A.3), however ERA40 estimates the tropical precipitation to be much higher than the other datasets. GPCP is the only dataset which is not reanalysis, and so is used to validate the precipitation biases. However the period which it covers is only half of the decadal hindcast period, so for validation beyond biases, NCEP was chosen, since firstly it covers the same period and secondly shows most similarity to the GPCP climatology.

Considering model biases, and focusing on the tropics where rainfall amounts are greatest; ECWMF (figures 4.2a - 4.2c) has a large wet bias over Indonesia (over 5mm/day) for all timescales. It is slightly too wet over the Amazon, and there is too much rain over the Gulf of Guinea in JJA, at the start of the West African monsoon. Over India in JJA there is a slight dry bias over the land to the west, whilst too much rain is falling over the ocean to the south of this - suggesting that the circulation pattern is not moving enough rain from the ocean to the land. The UKMO model (figures 4.2d - 4.2f) is generally too wet across the whole of the tropics, particularly over the oceans. The only large dry bias is over India during JJA, again suggesting poor monsoon simulation.

CERFACS (figures 4.2g - 4.2i) has a mixed wet/dry bias, where a dipole-like pattern emerges over several points in the ocean. This suggests that the rainfall amounts are correct, but are in the wrong place, likely due to erroneous circulation patterns caused by the large cold bias (seen in figures 4.1g - 4.1i). There is also a large wet bias over Southern Africa in DJF. IFM-GEOMAR (figures 4.2j - 4.2l) also has a dipole bias pattern over the ocean, suggesting wrongly-located precipitation. Generally the Amazon is too dry, as are Africa and India during JJA. However the bias over Africa is generally below 1mm/day throughout the year.

Looking at the bias of the multi-model ensemble (figures 4.2m - 4.2o), it can be seen that all models overestimate tropical precipitation, particularly over Indonesia. They also

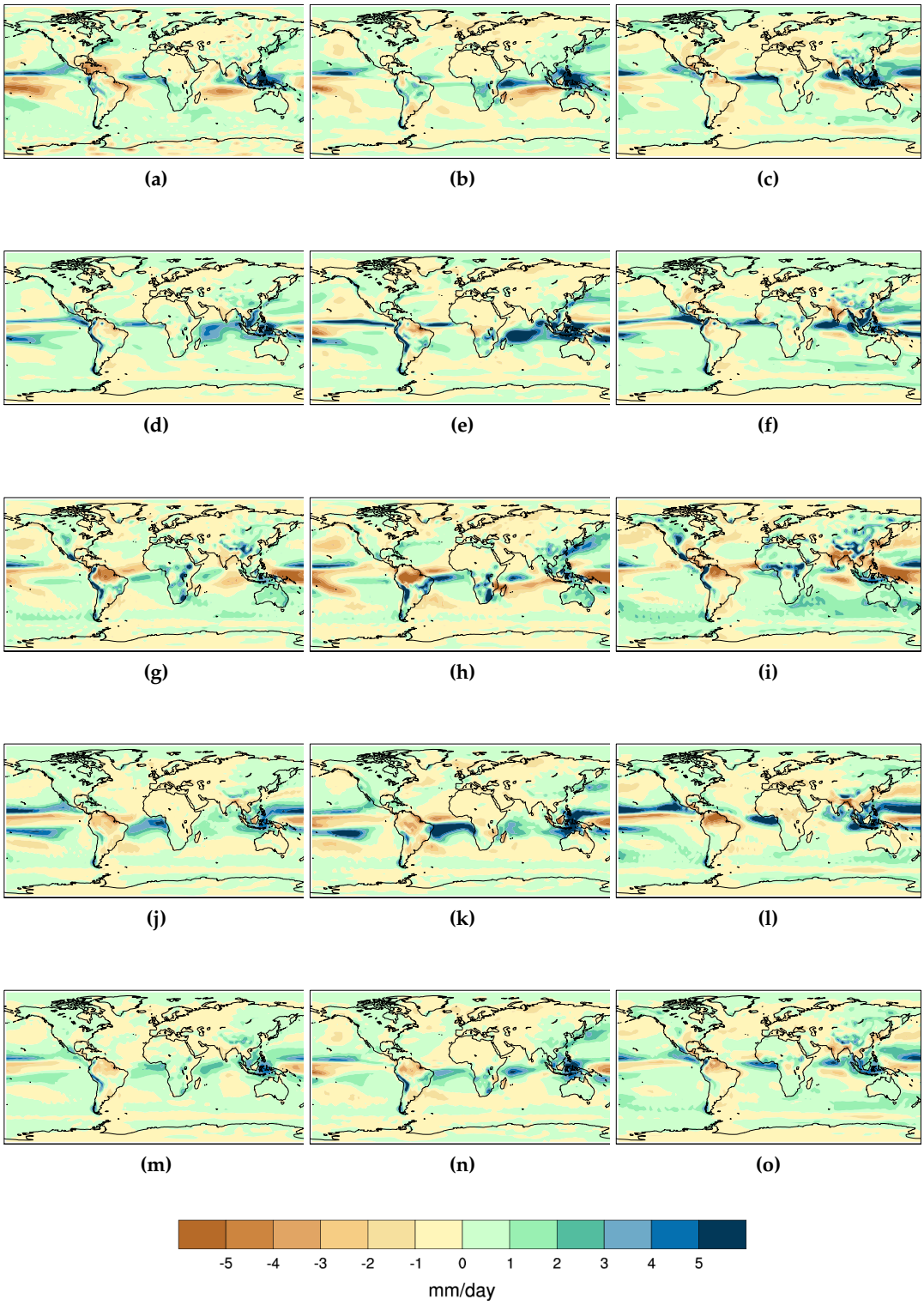


Figure 4.2: As figure 4.1, for precipitation. GPCP is used as a reference.

overestimate DJF rainfall over southern Africa and on average underestimate precipitation over the Amazon and over India.

Common biases suggest common problems to climate models, which, despite their differences share similar components. This is a drawback to the multi-model approach, since it cannot be said that the use of multiple models fully samples model uncertainty; at best only a subset of the uncertainty space is being explored. The presence (or absence) of significant biases does not necessarily mean that model predictions will have no skill. However it does suggest unrealistic dynamics, atmosphere-ocean connections and teleconnections, which may impact negatively on model predictions. It is to this skill the focus shifts, looking firstly at the correlations of year-to-year variations with observations.

4.3.2 Correlation

Figure 4.3 shows temperature and precipitation correlations between the ensemble mean of the hindcasts and the NCEP reanalysis before and after detrending, for annual, DJF and JJA averages. Before detrending, the temperature correlations are generally significant globally at the 99% level, with correlations reaching around 0.6. Generally the hindcasts have slightly larger correlations when validated against NCEP than against ERA40 (though the spatial patterns are similar). Over land, the regions which have significant correlations at the annual level for both reference datasets are located around equatorial Africa, the Mediterranean region, Asia, South-West US, and Greenland, with large areas of significant correlations over the north Atlantic and the Indian Ocean, with correlations of up to 0.6 (figure 4.3a). For seasonal averages the area of significant correlations is decreased, though with significant correlations across much of the globe, and maxima around the tropics (figures 4.3b and 4.3c). After detrending the temperature data (figures 4.3d, 4.3e and 4.3f), the correlations are not significant. This suggests that the significant correlations before detrending are caused by the long-term trend in temperature and beyond this any predictions of individual yearly or seasonal average temperature do not have skill.

Considering precipitation correlations (figures 4.3g, 4.3h and 4.3i), whilst there are areas of significance for NCEP, these lie generally over the ocean. Interest here is in the potential for disease prediction which occurs over land. Considering just the land grid points, correlations of precipitation are below significance almost everywhere. Furthermore, there are almost no areas of significance for the other references used (figure A.7), which suggests that the significant precipitation correlations shown in figure 4.3 are not robust. Therefore the ensemble mean precipitation of the models does not correlate with reality.

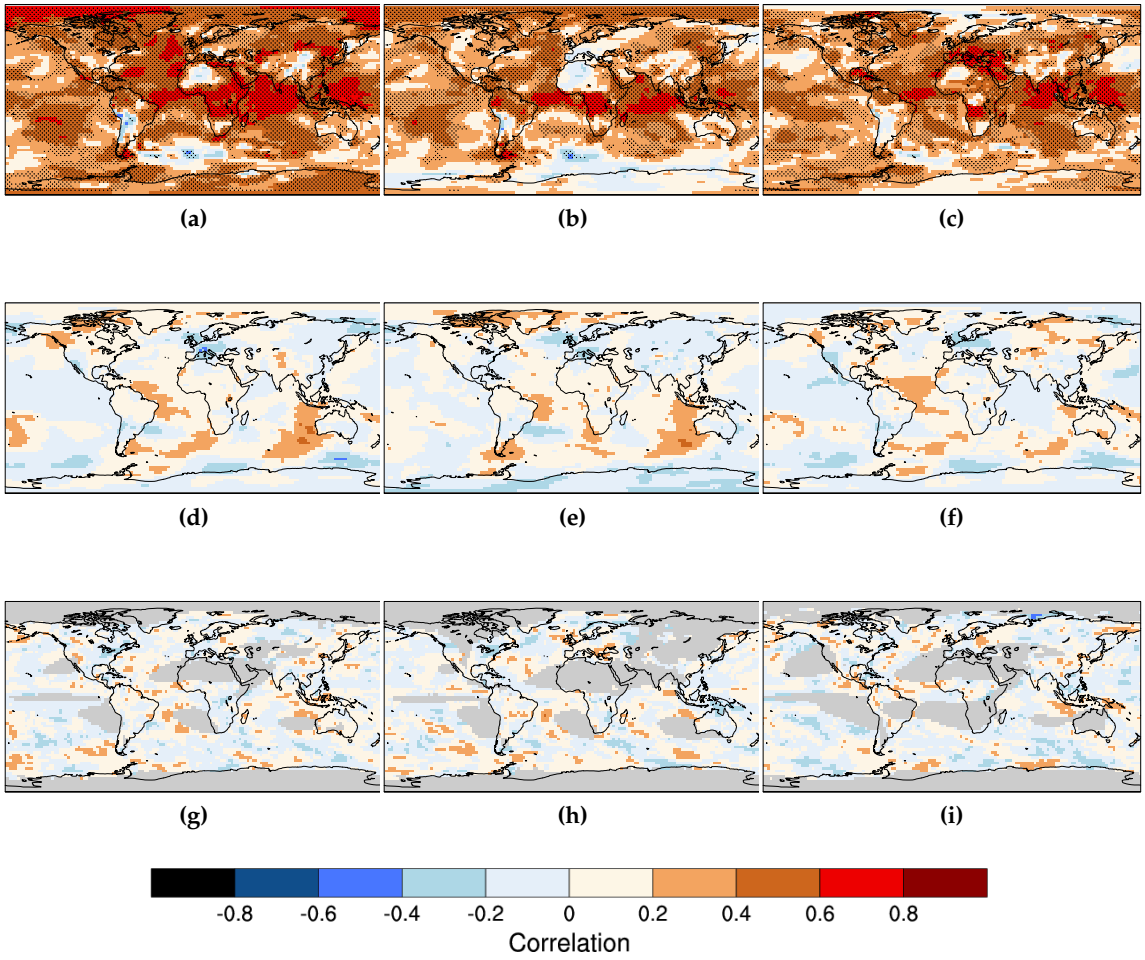


Figure 4.3: Correlation between the multimodel ensemble mean of the ENSEMBLES decadal hindcasts and NCEP reanalysis, calculated across all lead times and start dates. Temperature correlations before (after) detrending are shown in the top (middle) row; precipitation correlations are shown in the bottom row. Precipitation data has not been detrended before correlating. Results for annual, DJF and JJA averages (for all lead times) are shown in the left, centre and right columns respectively. The stipple area represents areas of correlation significant at the 99% level and the greyed out area in the precipitation plots indicates regions where model precipitation climatology is less than 1mm/day.

4.3.3 Potential predictability

Shown in figure 4.4 are maps of PP for temperature and precipitation. For temperature at the annual scale (figure 4.4a), there is a band of PP greater than 30% laying over the tropics, with a maximum of 60% over the maritime continent, with PP lower than 30% elsewhere. For seasonal averages (figures 4.4b and 4.4c) the PP maximum still lies over the tropics, though is generally not greater than 40%. This is consistent with studies showing that there is more predictability for temperature over the tropics than the extra-tropics (Palmer and Hagedorn, 2006a). There is also a slight seasonal pattern to PP, with a northward (southward) movement of the maximum PP band in boreal summer (winter), though this shift is not pronounced.

Precipitation PP is much lower than for temperature, which is also consistent with previous studies (Palmer and Hagedorn, 2006b). For annual averages over land it barely reaches greater than 5% (figure 4.4e). The maximum over the whole globe is 8%, over the Maritime Continent. This suggests that the spread of precipitation hindcasts over land is such that the potential for predictability may not be sufficiently high to be useful.

To demonstrate explicitly what the difference between high and low PP looks like, in figure 4.5 timeseries of the multi-model spread is plotted explicitly. The range of the 5-95% spread for every year is shown for temperature and precipitation, from three single grid points for each. These were chosen simply because of their values of PP, representing relatively low, medium and high values of PP for each variable separately (latitude, longitude points as well as the specific PP percentage value are shown inlaid on the plots).

Looking at a point with low PP (e.g. figure 4.5b) and comparing the ensemble for one year and another, one can see that the difference is minimal. This would suggest that the forecast system can not provide sufficient information to update our prior belief; this is also before considering whether the ensemble captures the reality. Conversely where the PP is high (e.g. figure 4.5e), one ensemble can often be clearly distinguished from another, so the prediction is potentially useful (as long as reality tends to lie inside the ensemble spread). One might use a threshold of PP as a minimum measure of necessary but not sufficient usefulness to a decision maker. That is, one might require a certain level of signal-noise of a forecast system for it to be useful. However the definition of this threshold depends on the context of the decision, taking into account individual risk-aversion, costs and losses. An arbitrary threshold cannot be chosen one-for-all, but at the least one can say that the higher the PP, the more potential use the forecast has (everything else being equal).

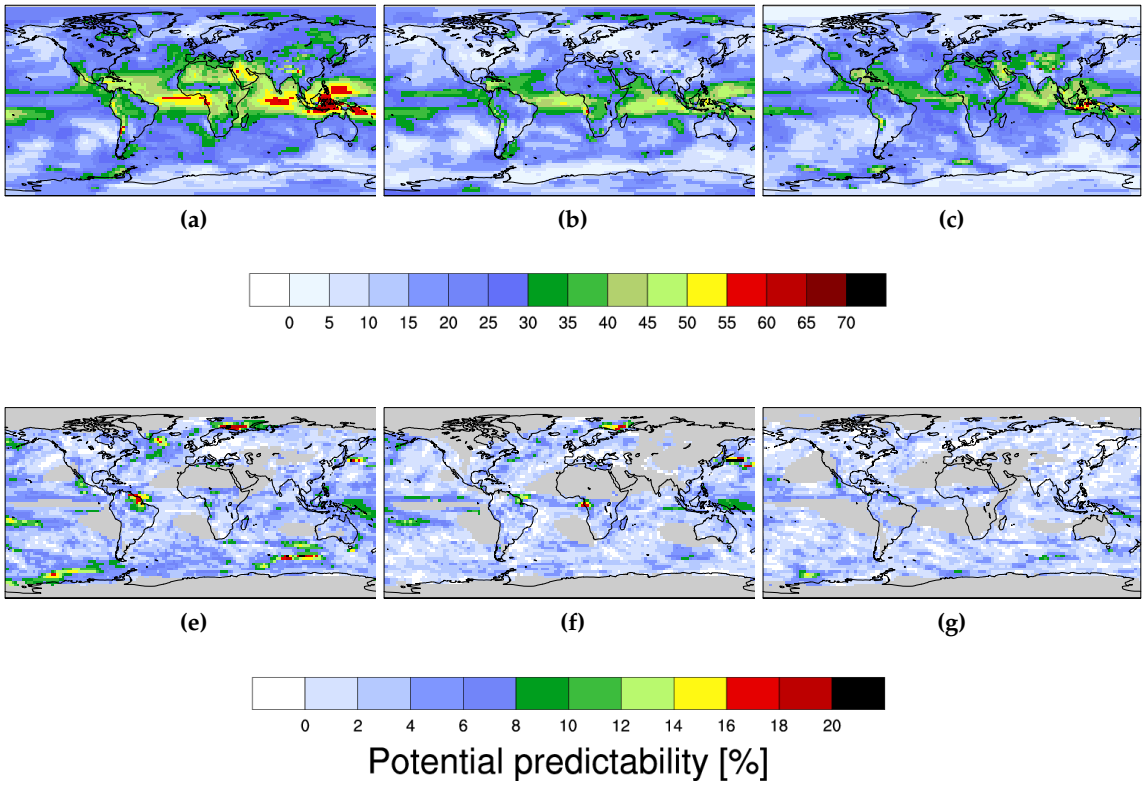


Figure 4.4: Temperature (left column) and precipitation (right column) potential predictability for annual averages (a & b), DJF (c & d) and JJA (e & f). Potential predictability is calculated across the concatenated timeseries of all lead times and all start dates. Greyed out area in the precipitation plots indicate regions where the ensemble mean climatology averaged across all models is less than 1mm/day. Note the different scale for the colour bar in the temperature and precipitation plots.

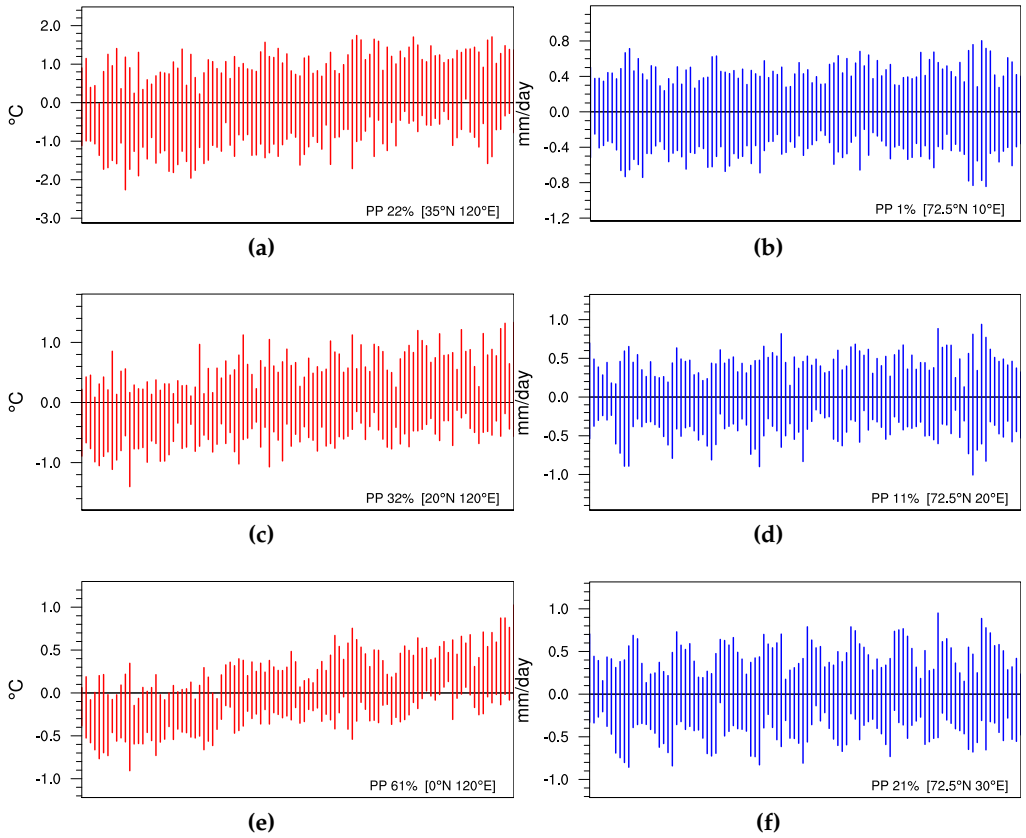


Figure 4.5: Time series examples of low, medium and high potential predictability for temperature (left column) and precipitation (right column), from single grid points.N.B. There is no scale on the x-axis; the scale is discontinuous as hindcasts have been concatenated and only the spread of each individual ensemble relative to the variability between years is important here - the coincidence of an ensemble with a particular year is irrelevant. Note the difference in ensembles between years, comparing the high and low PP (e.g. figure 4.5f and 4.5b); with lower PP, the ensembles are less distinguishable.

4.3.4 Multi-year trend correlation

Looking now at predictions for spatial averages, the correlation of multi-year trends is considered for several regions of the globe. These regions are shown in figure 4.6.

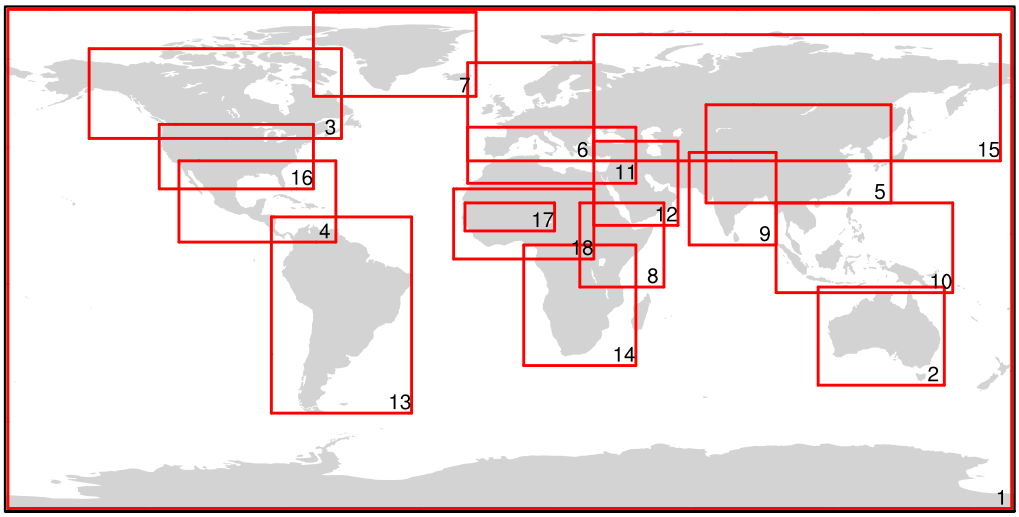


Figure 4.6: Regions considered for trend correlation analysis.

Figure 4.7 shows the significance levels for temperature and precipitation trend correlations for annual averages. For temperature at the annual scale, significant correlations for all models are observed for global land-sea and land only averages (figure 4.7a). This result is robust across both reference datasets (figure A.8). At regional scales (smaller than global) most regions have correlations below significance, however there are some regions where significant correlations are consistent across models and reference datasets. These are: Canada (UK Met Office/ECMWF, 7-9 year length), Central America (CERFACS, 5-7 year), China (CERFACS, 7-9 year), Horn of Africa (UK Met Office, 5-6 year), Mediterranean (ECMWF, 9-10 year), Middle East (UK Met Office/ECMWF/CERFACS, 5-6 year) and USA (ECMWF, 7-8 year).

For multi-year trends in seasonal averages, the correlations are less significant, suggesting that predictions for the smaller temporal averages are less skilful. Whilst there is some significance in global seasonal trends for NCEP, this does not hold for ERA40, where the number of points of significance is low.

Considering precipitation, figure 4.7b shows trend correlation significance levels for annual averages. These results for precipitation are not significantly better than what might arise from chance, and the pattern of significant points is not robust across different reference datasets (see figure A.9). This suggests that direct predictions of precipitation trends do not have skill for any region of the globe defined in this study.

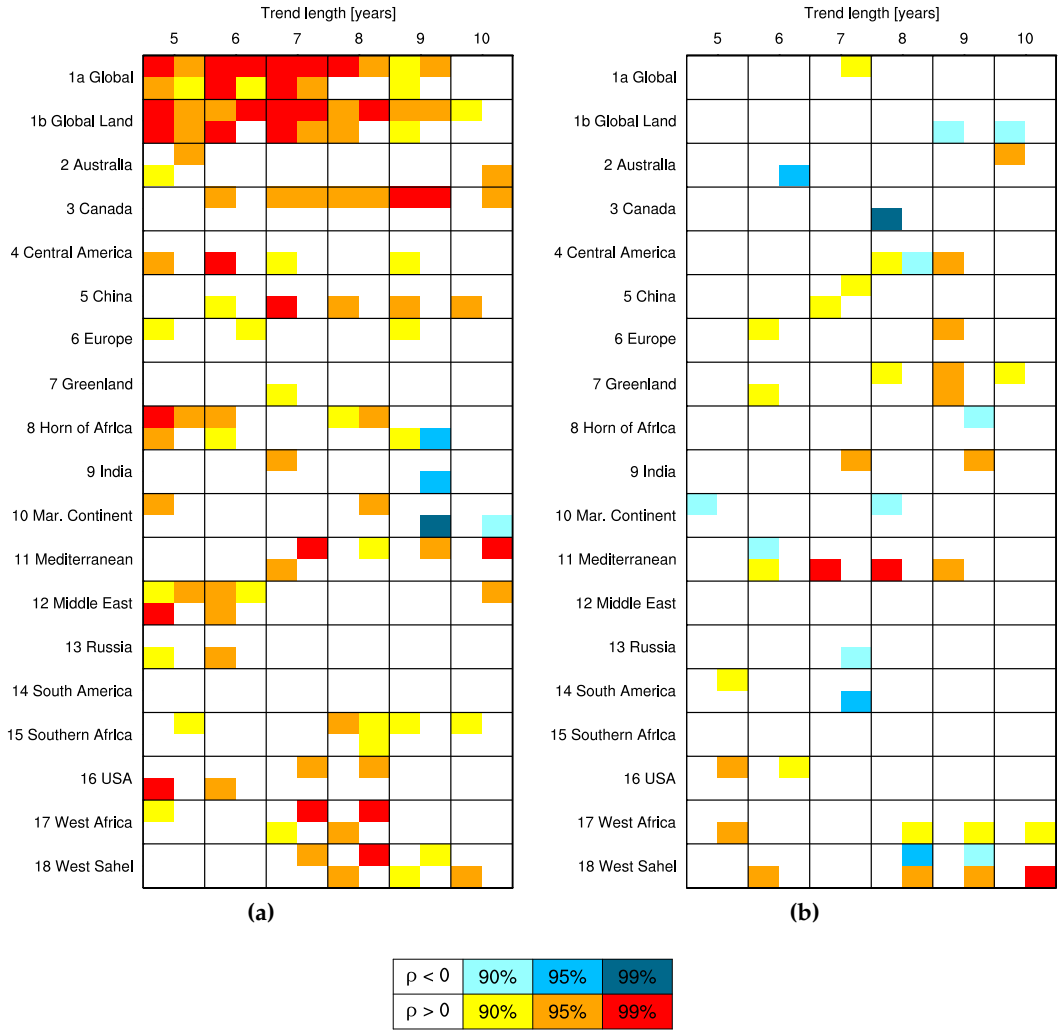


Figure 4.7: Multi year trend correlation significance levels for annually averaged temperature (left) and precipitation (right). The reference dataset for both is the NCEP reanalysis. Each quadrant in each square stands for one of the four models in the ENSEMBLES decadal hindcasts (clockwise from top left: UK Met Office, ECMWF, IFM-GEOMAR, CERFACS). The three variations in warm (cold) colours indicate Spearman's rank correlation coefficients significantly above (below) zero at the 90%/95%/99% levels respectively (levels at ± 0.324, ± 0.382, ± 0.491).

4.4 Discussion

The analysis of the ENSEMBLES decadal hindcasts presented here suggests that the prediction skill of the models for temperature is limited to annual global land-sea and global land trends, and that there is no skill in precipitation predictions. There are some areas of significant correlation over the globe for temperature before detrending, though with trends removed the correlation between hindcasts and multiple reference datasets are below useful levels of significance.

It has also been shown that the correlations and PP are lower for seasonal averages than for annual averages (figures 4.3 and 4.4). Trend correlations are significant for globally averaged temperature for multiple trend lengths (figure 4.7a), and a few regions have been identified which have significant temperature correlations, independent of the reference dataset. Correlations for precipitation are only above significance anywhere for NCEP, and for the other reference datasets they are below significance everywhere, suggesting that they are not robustly significant. Precipitation also has low PP (e.g. figure 4.4e), and precipitation trend correlations are below significance (figure 4.7b). These results are consistent with other studies, which find the global average temperature trend represented well in these hindcasts, and do not find significant skill for precipitation for four year averages (Oldenborgh et al., 2012). This work then extends this conclusion to predictions at annual and seasonal scales.

Whilst there may be skill in prediction of near term evolution of global average temperature, this does not equate to the ability to say something useful about climate on the scale on which it impacts human society (Oreskes et al., 2010). It is questionable how useful a skilful prediction of the annual global average temperature trend is if one is interested in making predictions of phenomena which unfold on regional, sub-annual scales. Annual global average temperature masks temporal and spatial variability, and it is variability which drives climate impacts (Washington et al., 2006). Coupled with poor predictions of precipitation, this suggests that current decadal predictions are not skilful at impact scales and that forecasts made by them are not of sufficient quality to drive models of climate impacts (such as agricultural, health or hydrological models).

Even though there may be little useful skill in direct model output of temperature and precipitation it may still be possible to predict impacts on decadal timescales. This may be done by relating impacts to potentially predictable variables other than air temperature and precipitation (Smith et al., 2010). It may also be possible to predict climate impacts using dynamical or statistical methods which predict the evolution of low frequency oceanic oscillations such as the Atlantic Multidecadal Oscillation and the Pacific Decadal Oscillation (e.g. Enfield and Cid-Serrano, 2006) and then relating them

to climate impacts. However, there are several potential sources of uncertainty to consider when attempting to predict impacts in this indirect way: uncertainty in the exact nature of the teleconnection between a large scale climate mode and regional climate impact, uncertainty in the prediction of the oscillation itself, and uncertainty due to the unpredictability of the forcing, due to the state of higher-frequency climate modes such as ENSO which are large source of climate variability and may potentially interact with low-frequency modes in unknown ways.

There are some limitations to these conclusions. The first is the limited number of start dates available in the hindcasts. Generally more validation points gives more confidence, and with only nine hindcast start dates confidence in validation is limited. A second limitation is the relatively small size of the initial condition ensemble. Ensemble size is particularly important when estimating potential predictability, and it is questionable that three members per model is sufficient to robustly estimate this. A final caveat to note is that there is uncertainty in the reference datasets to which the hindcasts are compared, particularly for reanalysis datasets, which are not necessarily representative of reality. This is particularly the case in places where observations are sparse (e.g. in Africa in the 1990s).

Finally, whilst the skill of the models is low, decadal prediction is still in its infancy. Furthermore, initialised climate models run in decadal mode have the potential to be useful in other ways. They have the advantage over uninitialized climate projections in that they can help to inform model development, are useful to learn about model biases, initialization strategies and climate variability. They can also help to build trust in climate projections. Nevertheless, the conclusion here is that the generation of decadal climate models used as part of the ENSEMBLES project have not demonstrated the ability to make useful predictions of climate at the scales at which it impacts society. This is a negative result, but it is nonetheless an important one for relevant communities to understand, as understanding current limitations of predictions allows efforts made to adapt to changes in climate to be focused wisely (Oreskes et al., 2010). Decadal models will continue to be developed (for example in the current CMIP5 experiments, WCRP, 2013), and their eventual role in climate change adaptation policy it is not yet clear. It is therefore important to maintain effective communication between research communities along the science-policy spectrum, such that research and policy expectations of decadal prediction remain informed by reality.

This is then the end of consideration of predictions on decadal timescales for the rest of this thesis, since they have been shown here to have insufficient skill to make useful predictions of climate-driven disease risk. Instead the focus shifts to shorter timescales, specifically seasonal forecasts. The next chapter then turns to the validation of seasonal

climate model hindcasts.

CHAPTER 5

An evolution of seasonal climate forecasting skill: comparing DEMETER, ENSEMBLES and System 4.

This chapter contains a comparison of the skill of seasonal climate forecasting systems. The systems studied are a set of hindcasts from the DEMETER project, the ENSEMBLES project and most recently a hindcast set from ECWMF's current seasonal model, System 4 (produced in 2004, 2008 and 2011 respectively).

5.1 Methodology

Seasonal predictability is discussed in section 2.1.2. Observational datasets and mathematics relating to skill scores are described in chapter 3. This methodology section describes the climate models used and forecast targets, and summarises the metrics used to compare the forecast systems.

5.1.1 Modelling systems

A summary of the seasonal forecast systems can be found in table 5.1. They are the DEMETER multimodel system, produced around 2004, the ENSEMBLES system produced around 2008 and the latest version of the European Centre for Medium Range

System name	Date	Period	# models	# members	Reference
DEMETER	2004	1980-2001	7	63	Palmer et al., 2004
ENSEMBLES	2008	1960-2005	5	45	Van Der Linden and Mitchell, 2009
ECMWF: System 4	2011	1981-2011	1	15	S4, 2013

Table 5.1: Summary of the seasonal hindcasts used in this study

DEMETER	ENSEMBLES
ECMWF	ECMWF
Météo-France	Météo-France
LODYC	UK Met Office
UK Met Office	IFM-GEOMAR
MPI	CMCC-INGV
CERFACS	
INGV	

Table 5.2: Origin of the models used in DEMETER and ENSEMBLES

Weather Forecast's (ECMWF) seasonal forecasting model, System 4, using the hindcast set produced in 2011.

The DEMETER seasonal re-forecasts consist of output from a multi-model ensemble of seven different fully coupled Atmosphere-Ocean Global Climate Models (AOGCMs), each run with nine different sets of initial conditions, whilst the ENSEMBLES multi-model ensemble comprises five AOGCMs, each with nine sets of initial conditions. The institutions providing each of the models for the projects are listed in table 5.2. Details of model resolution, atmospheric components and initialisation strategies for each of the DEMETER and ENSEMBLES models can be found in the references in table 5.1. Each of the models in DEMETER and ENSEMBLES was initialized four times per hindcast year, in February, May, August and November, and forecasts were made in each case six and seven months ahead of the start date respectively[1].

The System 4 hindcasts are based on the latest version of the operational coupled ocean-atmosphere model developed at ECMWF. The atmospheric model version is the IFS model at higher spatial resolution (about 0.7° x 0.7°) than the former system System 3, with a higher top of the atmosphere (0.01hPa) and more vertical levels (91). System 4 uses NEMO instead of HOPE as its ocean component, with initial conditions generated

[1]Fourteen month integrations were also carried out as part of the ENSEMBLES seasonal experiments; these are not considered here.

by the Near Real Time NEMOVAR suite instead of HOPE/OI. A fifteen member ensemble is created for each start date, with one start date at the first of each month; each ensemble member is simulated forward in time for seven months from the start date. Initial perturbations are defined with a combination of atmospheric singular vectors and an ensemble of ocean analyses. Atmosphere model uncertainties are simulated using the 3-time level stochastically perturbed parameterized tendency scheme and the stochastic back-scatter scheme (S4, 2013). Initial conditions come from ERA-Interim.

5.1.2 Forecast targets

For each modelling system, prediction of average temperature and rainfall during the rainy season is considered for several regions. The rainy season is the time of year with most interannual climate variability, and for the purpose of disease prediction is the time when climate is a large factor in vector-parasite dynamics. Anomalous climate conditions during the rainy season are a large forcing on epidemiological outcomes, so the ability to predict climate at these times of the year is important if one wishes to make useful predictions of disease.

For regions where vector-borne diseases are prevalent, biases are firstly considered. Biases are relevant for impact models as they and often require bias correction of climate model input before their use. If biases are naturally low then bias correction techniques do not modify the input greatly. A driving model with a low bias is preferable for impact modelling, since post-processing techniques can introduce uncertainty, particularly for precipitation.

Forecast skill is then measured by looking at the correlation of ensemble means with observations, and for relative operating characteristic area under curve (ROC AUC) for upper and lower tercile events. Details of ROC AUC can be found in section 3.2. Note that when calculating tercile events for each of the multi-model systems, tercile thresholds and forecast probabilities were determined for each individual model separately before combining to create an ensemble probability forecast.

In each case the reference for temperature used is the NCEP reanalysis and for precipitation the GPCP dataset (see section 3.1 for further details). These were chosen as they cover the necessary hindcast period, and since GPCP is generally considered to be closer to reality than precipitation from reanalysis.

The regions and target seasons chosen for study are listed in table 5.3; for each region the seasonal targets have been defined by plotting the climatology from GPCP and selecting

Region	Target	Forecast start	Sub regions
West Africa	JAS	May	Sahel, Gulf of Guinea
Southern Africa	DJF	November	Malawi, Botswana
East Africa	MAM	February	Kenya
Indian subcontinent	JJA	May	West India, Bangladesh

Table 5.3: A summary of the seasonal forecast target regions and seasons.

the three month season of maximum rainfall. As there are only four start dates per year for DEMETER and ENSEMBLES, the choice of start date depends on the target. For example a forecast for June-August (JJA) must use the May start date. Since System 4 has one start date per month, the forecast chosen for comparison is the one initialized at the same date as the other forecast systems. Finally since each forecast system has a different hindcast period are different (see table 5.1), to ensure a fair comparison an overlapping subset of years is studied for each (that is, 1981-2001).

Subsequently for each large region, spatial averages of smaller regions have been studied, these regions are shown in figure 5.1. Metrics used to indicate the usefulness of the forecasts in these subregions are reliability maps, the Brier Skill Score (BSS) and its decomposition, and the economic value (see section 3.2 for further information the validation metrics).

These smaller regions were randomly selected without taking into account *a priori* where the models are skillful. The reason for this is that it is closer to the reality of working with an end-user. Whilst the motivation of a climate modeller may be to showcase the very best of their model, an end-user is likely to have an interest only in their local area; a Malawian farmer has little interest in a forecast which has no skill over Malawi but skill elsewhere. Certainly, organisations acting internationally will be interested in forecasts for many local regions, and the subregions chosen here are only a limited look at the skill of the forecasting systems. However in the choice of regions to study it is preferable to be as unbiased as possible: given the variability in skill, regions could be chosen to present any model of choice in a good or bad light. By selecting regions *a priori* this bias is eliminated.

A discussion of results for West and southern Africa and their associated subregions follows; the remaining figures and discussion for the Horn of Africa and the Indian Subcontinent and their subregions are contained in appendix B.

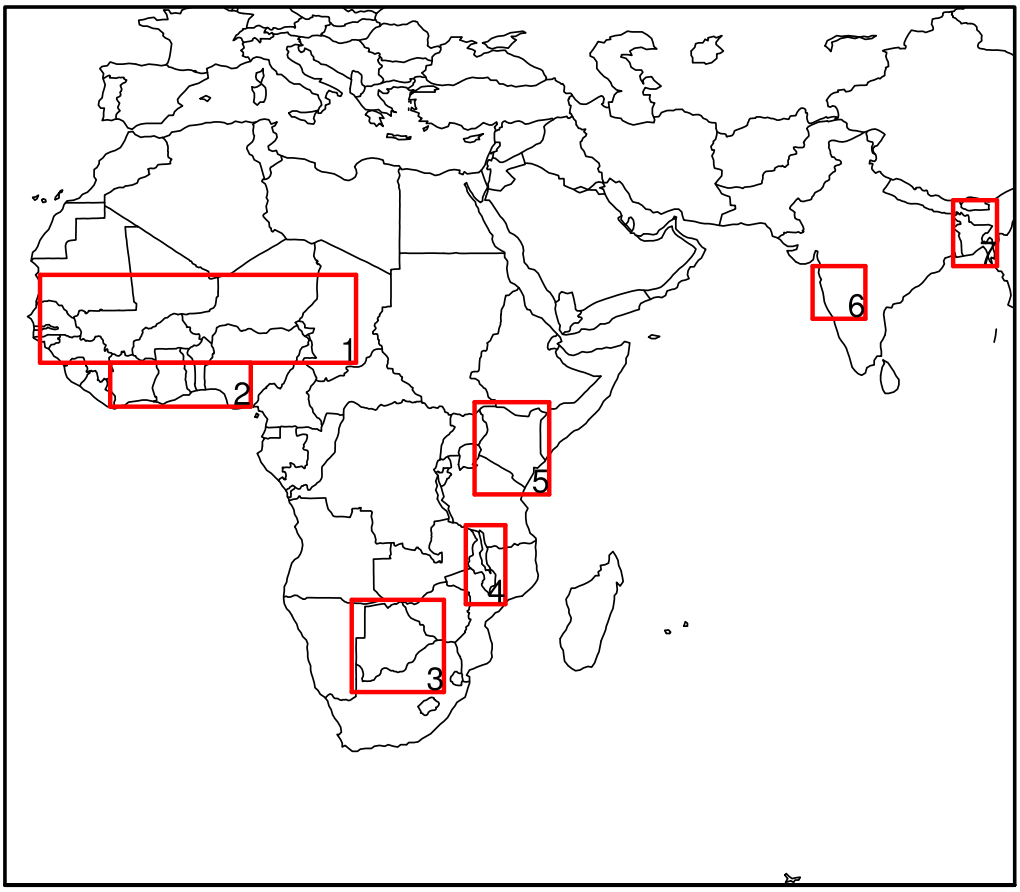


Figure 5.1: Sub-region definitions. 1-7: the Sahel, Gulf of Guinea, Botswana, Malawi, Kenya, West India and Bangladesh.

5.2 Results

All results comparing DEMETER, ENSEMBLES and System 4 are summarised in table 5.4. Subsequently results for each region are described in separate sections.

5.2.1 West Africa

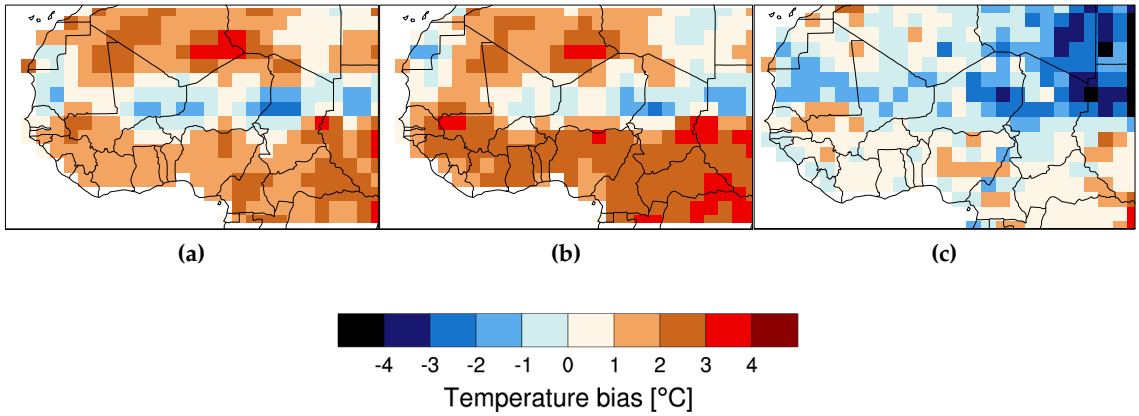


Figure 5.2: Ensemble mean JAS average temperature bias over West Africa vs NCEP, for DEMETER, ENSEMBLES and System 4 (a-c).

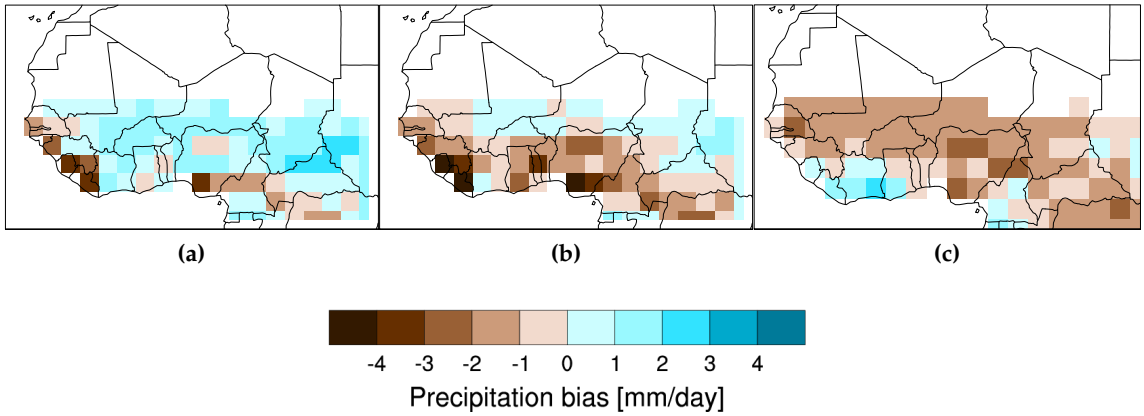


Figure 5.3: As figure 5.2, for ensemble mean precipitation vs GPCP. Only points are shown where the observed seasonal climatology is greater than 1mm/day.

Results are presented first for West Africa. In figure 5.2 the biases in temperature for the three systems are shown. The warm bias increases between DEMETER and ENSEMBLES, but System 4 shows a marked improvement; for most of the region where rainfall is present the bias is reduced DEMETER and ENSEMBLES show a similar pattern of bias, with a strip of cold bias around the northern Sahel.

	Temperature	Precipitation
West Africa		
Bias	Steady improvement between the systems	Wet bias in DEMETER reduced in ENSEMBLES. Bias is smallest near the coast in System 4
Correlation	Improvement between DEMETER and ENSEMBLES, no further increase in System 4	System 4 highest for most of the domain
ROC AUC	ENSEMBLES and System 4 improved from DEMETER	System 4 highest for most of the domain, though DEMETER highest over the Ivory Coast
Southern Africa		
Bias	Small change between DEMETER and ENSEMBLES; improvement in System 4	DEMETER and ENSEMBLES similar; large improvement in System 4
Correlation	Significant over most of the domain in DEMETER and ENSEMBLES; slight reduction in some areas in System 4	Slight improvement between DEMETER and ENSEMBLES; large improvement in System 4
ROC AUC	High for all systems, though best in DEMETER by a small margin	Below significance for DEMETER; ENSEMBLES and System 4 show skill for some regions
East Africa		
Bias	Large in DEMETER and ENSEMBLES; reduced in System 4	Smallest in System 4
Correlation	Increase between DEMETER and ENSEMBLES, slight reduction in System 4	Steady improvement between the systems
ROC AUC	No significant increase between the systems	Increase between DEMETER and ENSEMBLES, slight reduction in System 4
Indian Subcontinent		
Bias	Similar pattern of warm and cold bias in DEMETER and ENSEMBLES. Cold for most of the region in System 4	Similar complex pattern in DEMETER and ENSEMBLES; magnitude reduced in System 4
Correlation	Significant in some areas in all systems; ENSEMBLES significant over the largest area	Similar between all systems
ROC AUC	Similar between all systems: significant over west coast India and Myanmar	No skill for most of the region; some skill over Nepal for all systems and over the southern tip of India for System 4 alone

Table 5.4: Summary of temperature and precipitation biases, ensemble mean correlation and ROC AUC for DEMETER, ENSEMBLES and System 4. A summary of brier skill score and potential economic value can be found separately in table 5.5

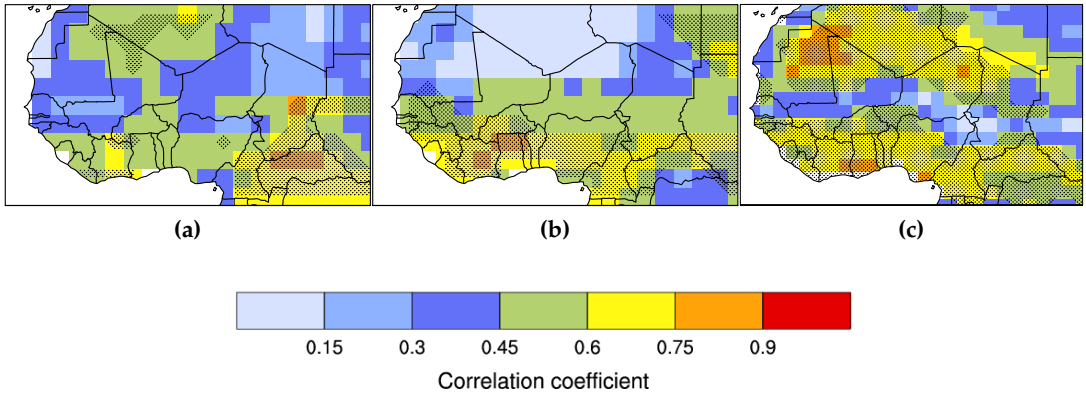


Figure 5.4: Pearson's product-moment correlations of JAS ensemble mean precipitation vs NCEP, for DEMETER, ENSEMBLES and System 4 (a-c). Forecasts issued at the start of May. Stippled area shows 99% confidence level, with sample size adjusted to take account for autocorrelation.

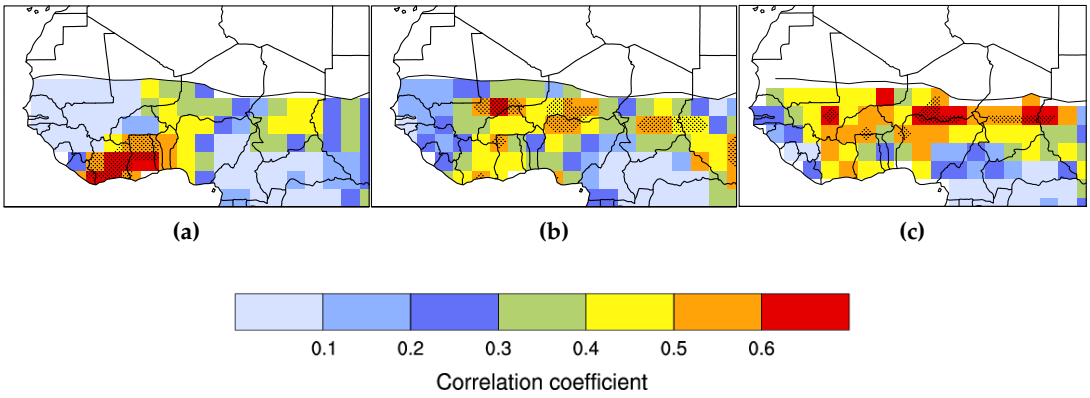


Figure 5.5: As figure 5.4, for ensemble mean precipitation vs GPCP. Only points are shown where the observed seasonal climatology is greater than 1mm/day.

For precipitation biases (shown in figure 5.3), DEMETER has a wet bias of a few mm/day across the Sahel, which is reduced in ENSEMBLES and in System 4 becomes a dry bias of over a mm/day. Close to the coast, there is no clear signal in DEMETER, but in ENSEMBLES the bias is coherently dry. This bias near the coast is reduced in System 4, but over the Sahel there is a slight dry bias.

For correlations of ensemble mean temperature (figure 5.4), there is a definite improvement between DEMETER and ENSEMBLES, with the most of the Gulf of Guinea below significance for DEMETER and above significance in ENSEMBLES. showing significant correlations. System 4 shows a similar level of correlation as ENSEMBLES near to the coast, though the skill in the north over the Sahara is much higher (how useful seasonal forecasts are in a desert however is questionable).

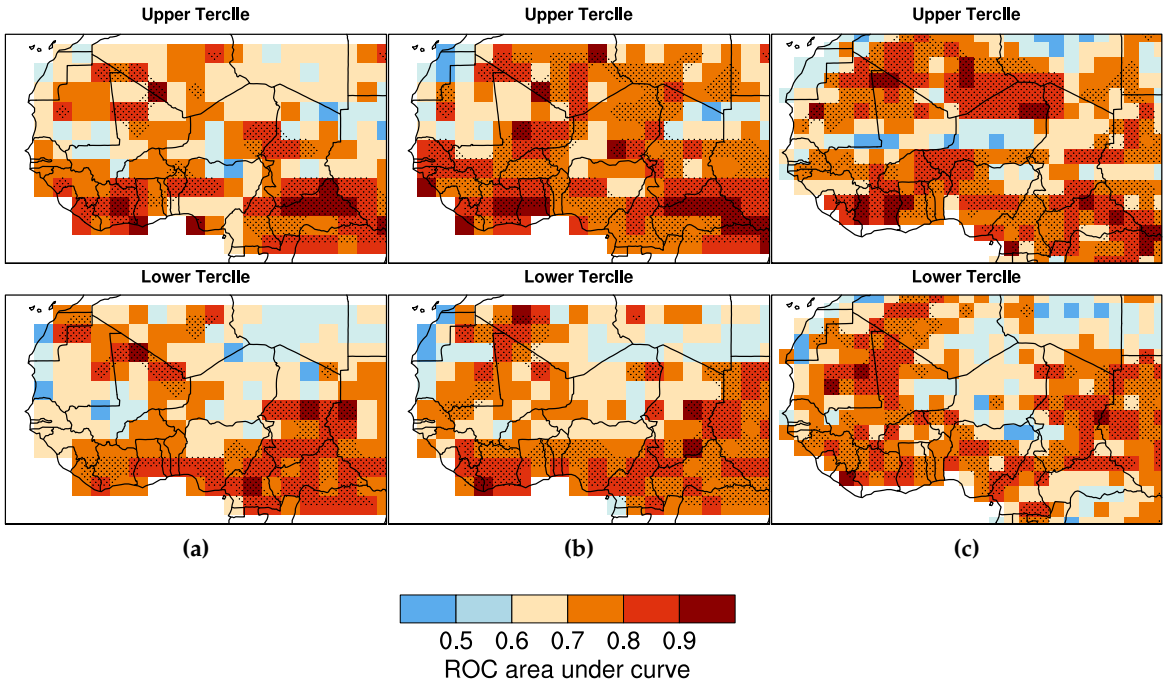


Figure 5.6: Relative operating characteristic area under curve (ROC AUC) for JAS temperature vs NCEP, for the May start dates of DEMETER, ENSEMBLES and System 4 (a-c). Stippled area indicates where the AUC is significant at the 95% level.

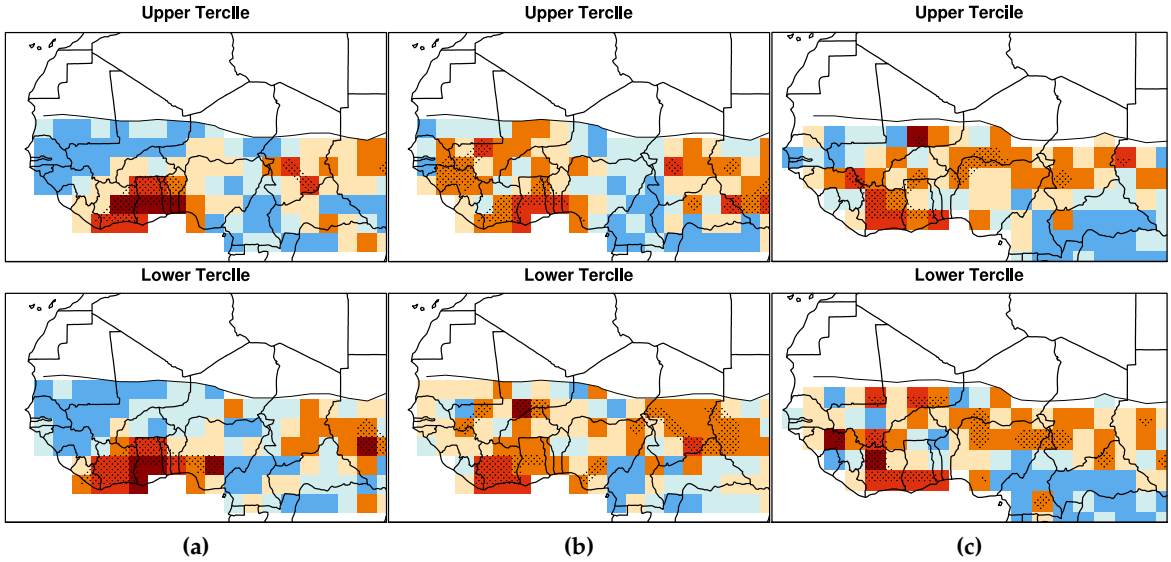


Figure 5.7: As figure 5.6, for precipitation vs GPCP. Only points are shown where the observed seasonal climatology is greater than 1mm/day.

Precipitation correlations (figure 5.5) show significance in a small region around the coast in DEMETER, near the Ivory Coast and Ghana. Elsewhere they are not significant. ENSEMBLES and System 4 do not have significant correlations near the coast as in DEMETER, though ENSEMBLES has some significant correlation further north around the Sahel. System 4 has the highest correlations for most of the domain, though the correlations above significance lie in a band around the limit of the >1mm/day region, stretching west to east for several contiguous grid points.

Plots of relative operating characteristic area under curve (ROC AUC) for temperature (figure 5.6) show that ENSEMBLES and System 4 have significant skill in the same regions as they show significant correlations (generally around the coast), with no noticeable differences in skill between them. Both show some improvement from DEMETER, particularly for upper tercile events around the coast of Nigeria where scores are above significance for System 4 and ENSEMBLES whilst being below significance in DEMETER.

For precipitation ROC AUC (figure 5.7), the areas where scores are significant are somewhat related to the pattern of significant correlations, though not precisely. For DEMETER the areas of significant ROC AUC and correlation match up almost exactly, with a high ROC AUC around the Ivory Coast/Ghana. However the significant correlations around the Sahel are not reflected completely in ENSEMBLES and System 4, whilst they both do have some significant patches of skill, the highest area for both is in the same region as DEMETER, around the coast. The level of ROC AUC here does not reach the same value as DEMETER, though outside of this region the score is generally higher in ENSEMBLES and System 4. Across the whole domain System 4 has the highest score, with most points taking values above 0.6, however these scores are not significant.

Looking now at the Sahel, figure 5.8 shows reliability plots for temperature, including the Brier skill score (BSS) and its decomposition. The slope of the reliability curve is positive for all systems, and there is a significant improvement in BSS between DEMETER and ENSEMBLES for upper tercile, increasing from 0.09 to 0.20. The BSS for lower tercile events has not improved. From the decomposition the improvement for upper tercile forecasts has mostly come from improvement in the resolution component. The sharpness has also improved, as can be seen from the differing distribution of the histograms. This improvement does not continue for System 4 upper tercile events, but the score for lower tercile is higher than both DEMETER and ENSEMBLES. The improvement in System 4 comes equally from improvement in reliability and in resolution; however for both upper and lower tercile forecasts the system over-predicts above forecast probabilities of 80%.

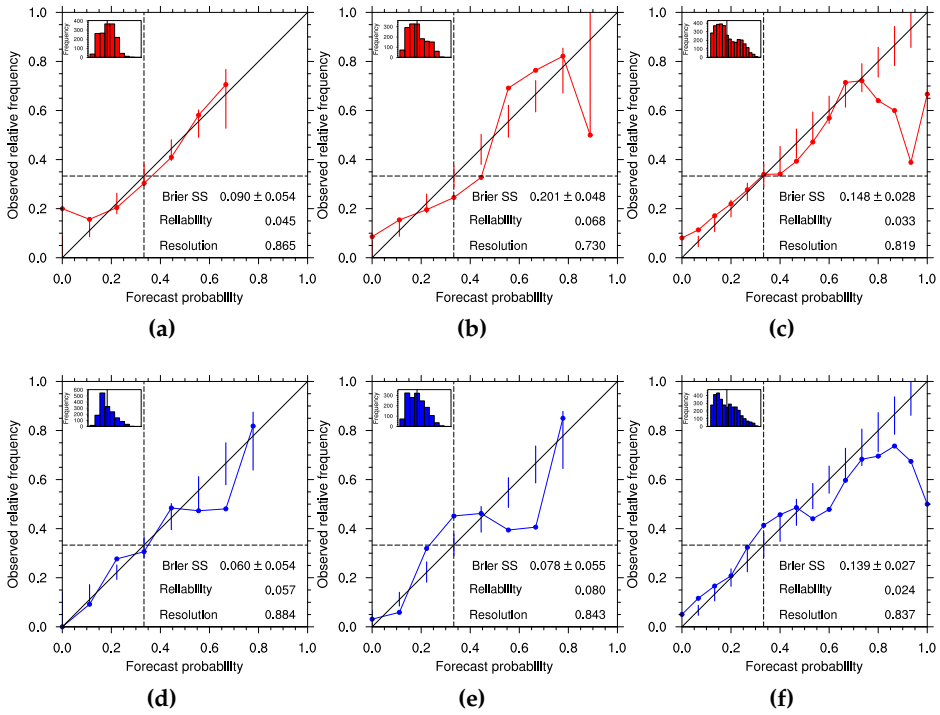


Figure 5.8: Reliability of upper (a-c) and lower (d-f) tercile JAS temperature forecasts over the Sahel (vs NCEP): DEMETER (a & d), ENSEMBLES (b & e) and System 4 (c & f). Forecasts issued at the start of May.

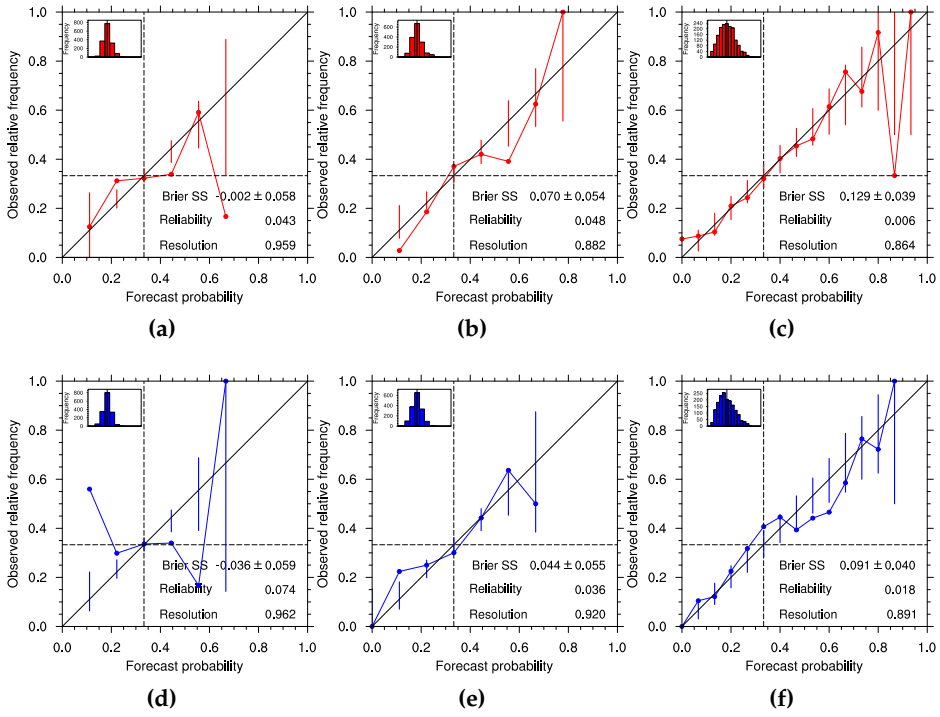


Figure 5.9: Reliability of Sahel precipitation vs GPCP, details as in figure 5.8.

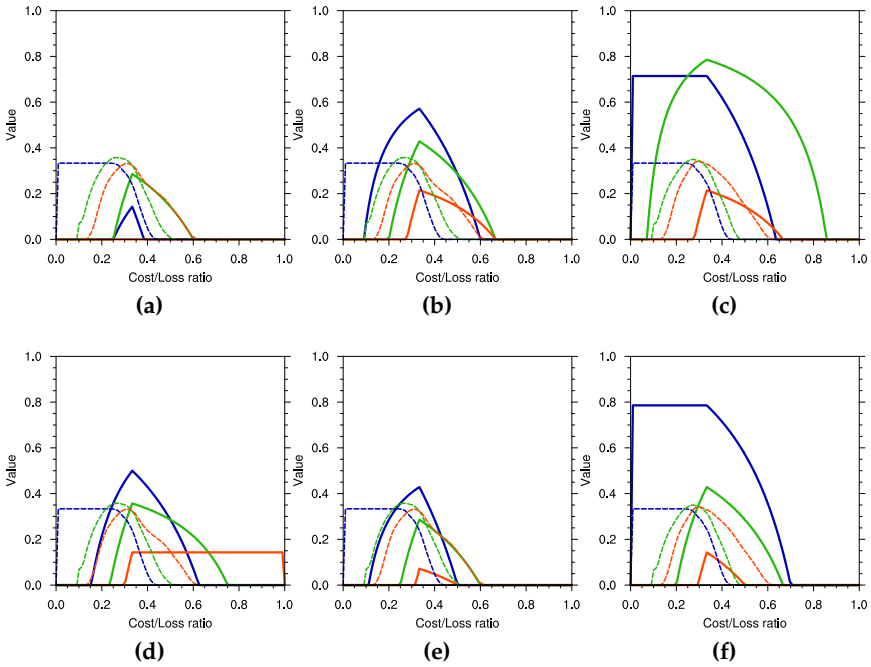


Figure 5.10: Value of upper (a-c) and lower (d-f) tercile JAS temperature forecasts over the Sahel, for DEMETER (a & d), ENSEMBLES (b & e) and System 4 (c & f). Curves for 30%, 50% and 70% decision thresholds are shown by blue, green and orange lines with dashed lines indicating 95% significance level for each threshold. Forecasts issued at the start of May.

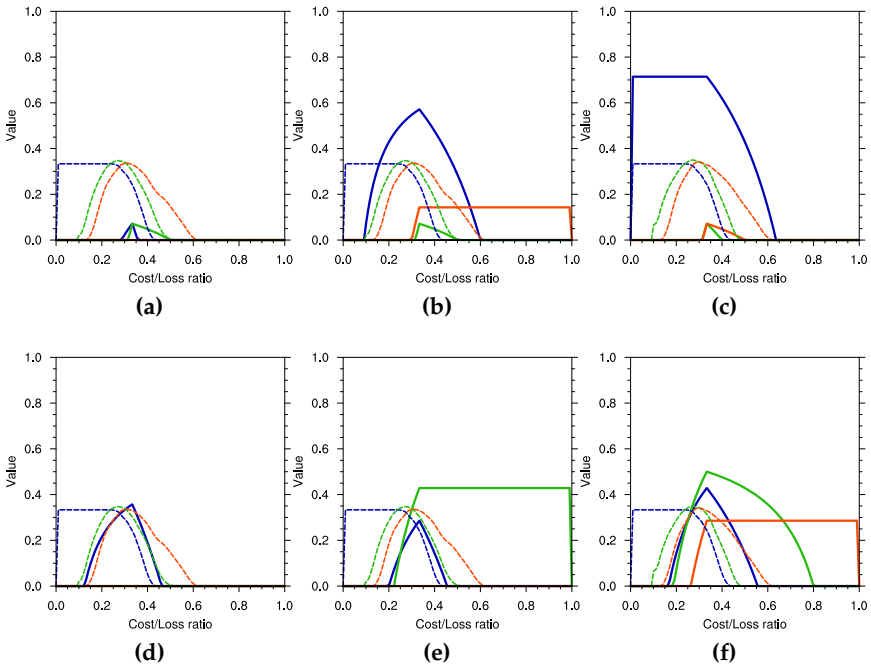


Figure 5.11: Value of Sahel precipitation vs GPCP, details as in figure 5.10.

For reliability of precipitation forecasts (figure 5.9), upper tercile has the highest BSS for System 4. This is due to a low value of the reliability component, as is reflected by how all points bar one lie inside the consistency bars. Lower tercile event forecasts for DEMETER and ENSEMBLES do not show as great an improvement over as they do for upper tercile events and improvement for these is slight. The BSS of System 4 can be said to be higher from the BSS for DEMETER (based on the standard error range), but the same cannot be said when System 4 is compared to ENSEMBLES.

The value curves of upper tercile temperature forecasts (figure 5.10) are below the 95% significance level for DEMETER. For ENSEMBLES they are 95% significance level. System 4 has decision thresholds which give a high positive value above the significance level, suggesting that there is a clear value to using these forecasts over climatology. For precipitation (figure 5.11), DEMETER forecasts for show no value for upper or lower tercile events, whilst both ENSEMBLES and System 4 show some positive value above significance at the 30% threshold for upper tercile events, and at mostly the 50% threshold for lower tercile events.

Turning to the Gulf of Guinea region and the reliability of temperature forecasts (figure 5.12), upper tercile forecasts show an improved BSS between DEMETER and ENSEMBLES, with no significant improvement beyond this in System 4. The reliability component for System 4 BSS is improved, but this is offset by a worsening of the resolution component. The comparison of lower tercile forecasts is similar; System 4 has the highest BSS, with a significant improvement in reliability and worsening of resolution. For precipitation (figure 5.13), the Brier Skill Score steadily worsens between DEMETER, ENSEMBLES and System 4, for upper and lower tercile forecasts. In both cases this is due to the resolution component steadily worsening and being offset slightly by an improving reliability component (as can be seen from the reliability curve). The sharpness of forecasts also increases, with System 4 forecasts occupying the entire domain of forecast probabilities.

Value of temperature forecasts (figure 5.14) is improved for upper tercile events dramatically between DEMETER, ENSEMBLES and System 4; DEMETER has no value above significance at any decision threshold, whilst ENSEMBLES has significant value at the 30 and 50% thresholds. System 4 has a large positive value at all three thresholds, particularly for the 30% threshold, which has value close to one over part of the cost/lost domain, indicating that it offered an almost perfect forecast of upper tercile events over the reference period. Lower tercile value is not as high for any system: DEMETER forecasts again have no value above significance, whilst for ENSEMBLES only the 30% threshold is above significance. The same is true for System 4, though the magnitude of value is not as high as ENSEMBLES and is only just above significance.

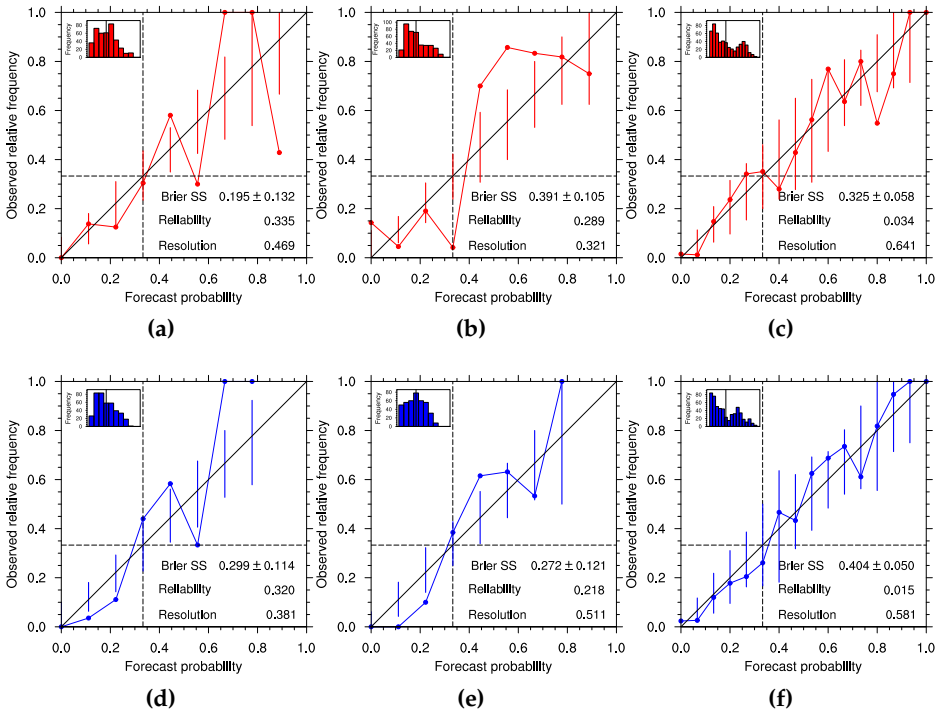


Figure 5.12: Reliability of upper (a-c) and lower (d-f) tercile JAS temperature forecasts over the Gulf of Guinea (vs NCEP): DEMETER (a & d), ENSEMBLES (b & e) and System 4 (c & f). Forecasts issued at the start of May.

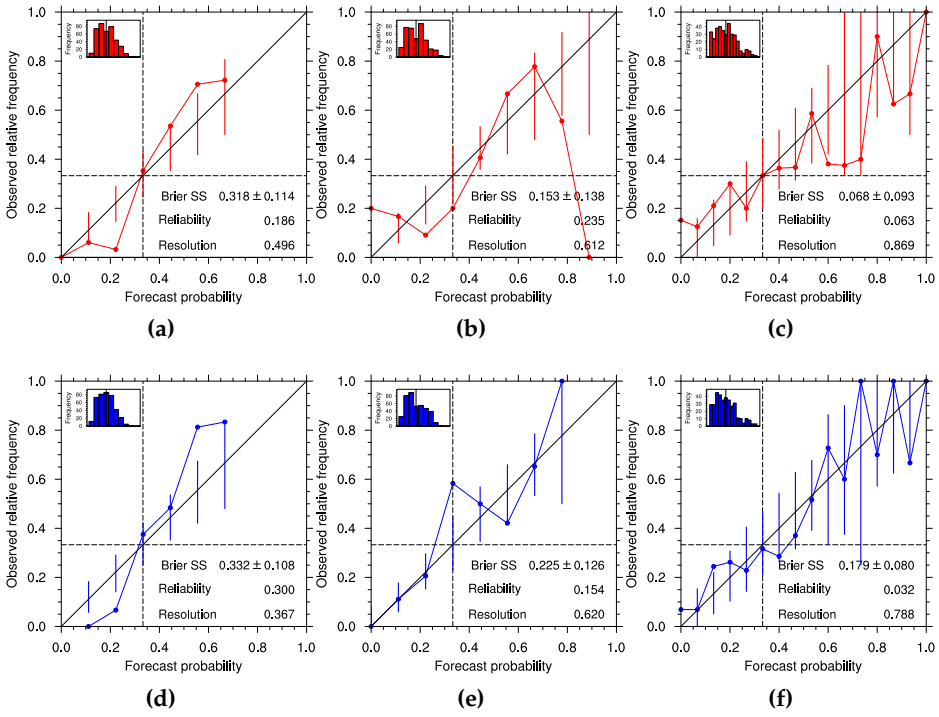


Figure 5.13: Reliability of Gulf of Guinea precipitation vs GPCP, details as in figure 5.12.

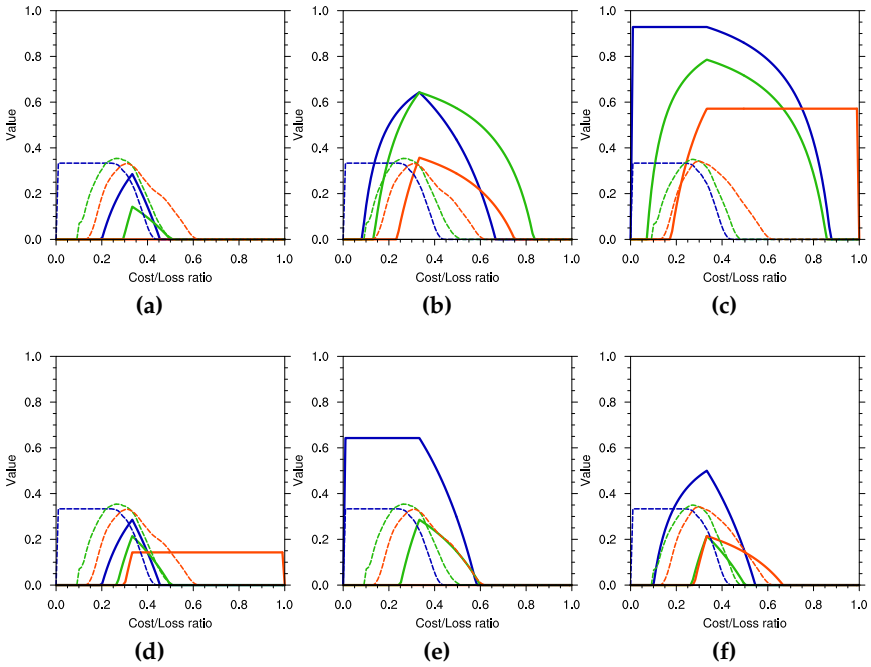


Figure 5.14: Value of upper (a-c) and lower (d-f) tercile JAS temperature forecasts over the Gulf of Guinea, for DEMETER (a & d), ENSEMBLES (b & e) and System 4 (c & f). Curves for 30%, 50% and 70% decision thresholds are shown by blue, green and orange lines with dashed lines indicating 95% significance level for each threshold. Forecasts issued at the start of May.

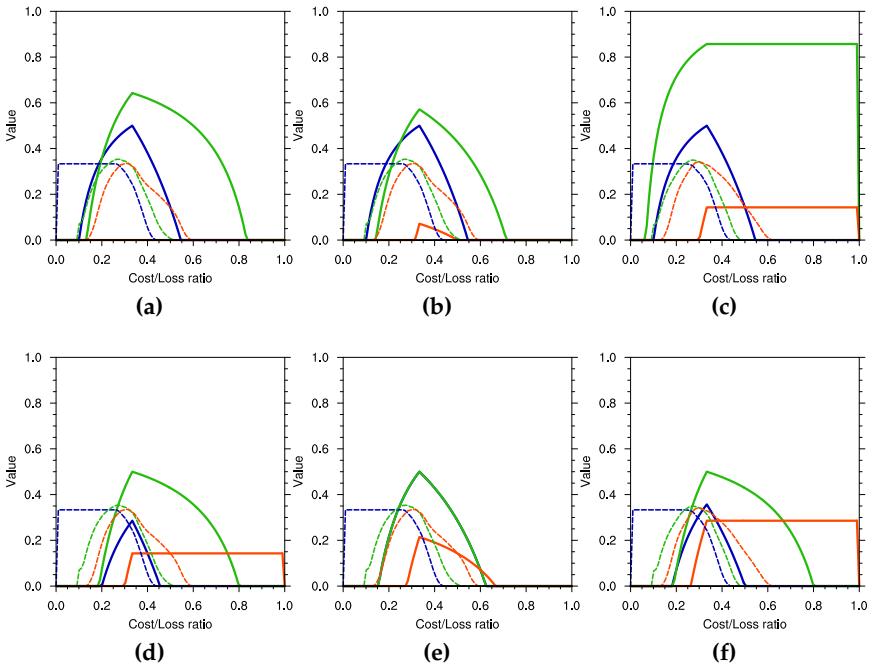


Figure 5.15: Value of Gulf of Guinea precipitation vs GPCP, details as in figure 5.14.

For precipitation (figure 5.15) all systems have value above significance for the 30 and 50% thresholds, with the highest value occurring using the 50% threshold in System 4. For lower tercile events the 50% threshold generally has the highest value above significance for all systems, though there is no improvement between DEMETER and System 4.

5.2.2 Southern Africa

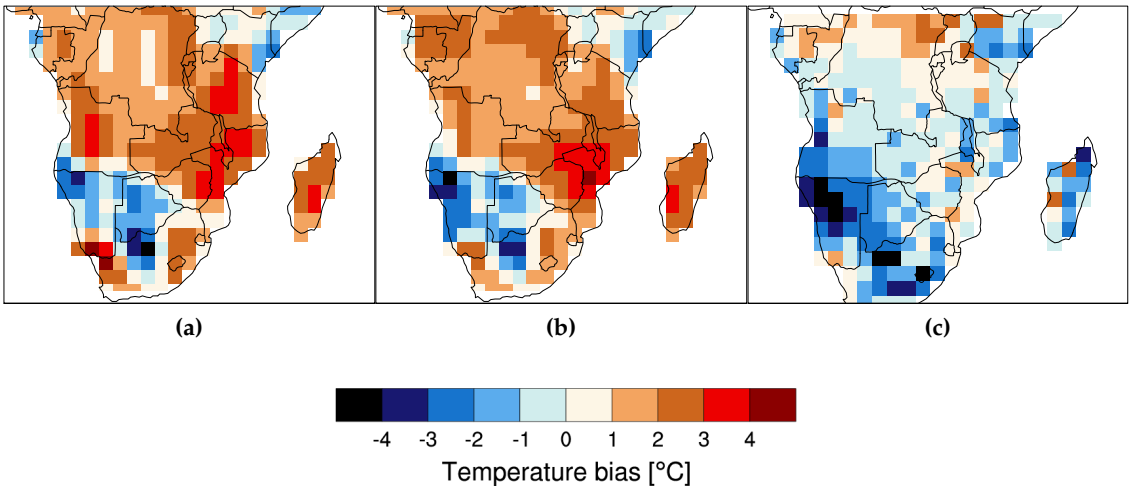


Figure 5.16: Ensemble mean DJF temperature bias vs NCEP, for DEMETER, ENSEMBLES and System 4 (a-c).

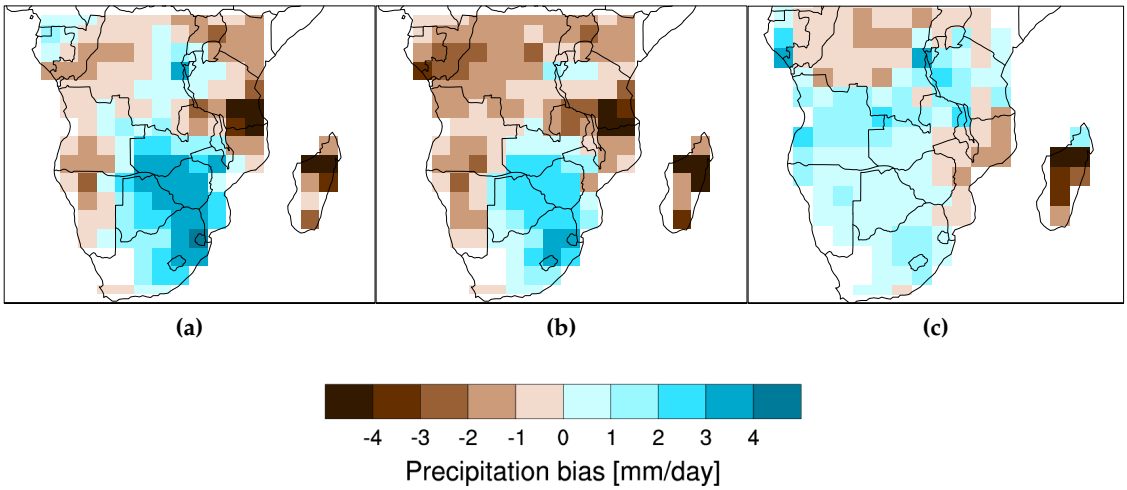


Figure 5.17: As figure 5.16, for ensemble mean precipitation vs GPCP. Only points are shown where the observed seasonal climatology is greater than 1mm/day.

Temperature biases for southern Africa are shown in figure 5.16. DEMETER and ENSEMBLES show a similar pattern, with not much improvement between the two. Both have a cold bias over the deserts in the south west. A warm bias surrounds this region, with a local maximum over Malawi. System 4 (figure 5.16c) also shares the cold bias over the desert with DEMETER and ENSEMBLES, but outside this region the bias is mostly much reduced; instead of a warm bias of over one and up to four degrees there is a slight cold bias of mostly under one degree. Madagascar has an

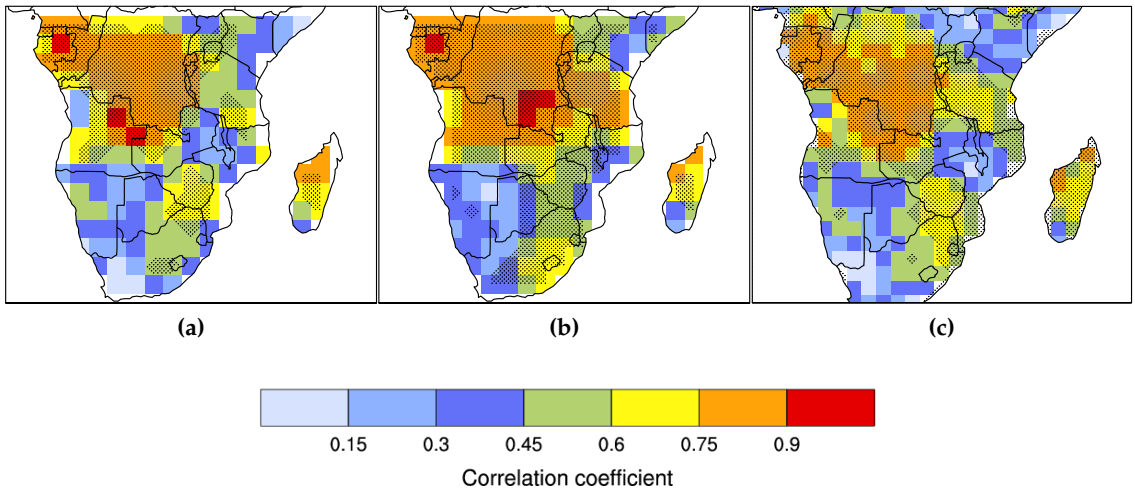


Figure 5.18: Pearson's product-moment correlation of DJF ensemble mean precipitation vs NCEP, for DEMETER, ENSEMBLES and System 4 (a-c). Forecasts issued at the start of November. Stippled area shows 99% confidence level, with sample size adjusted to take account for autocorrelation.

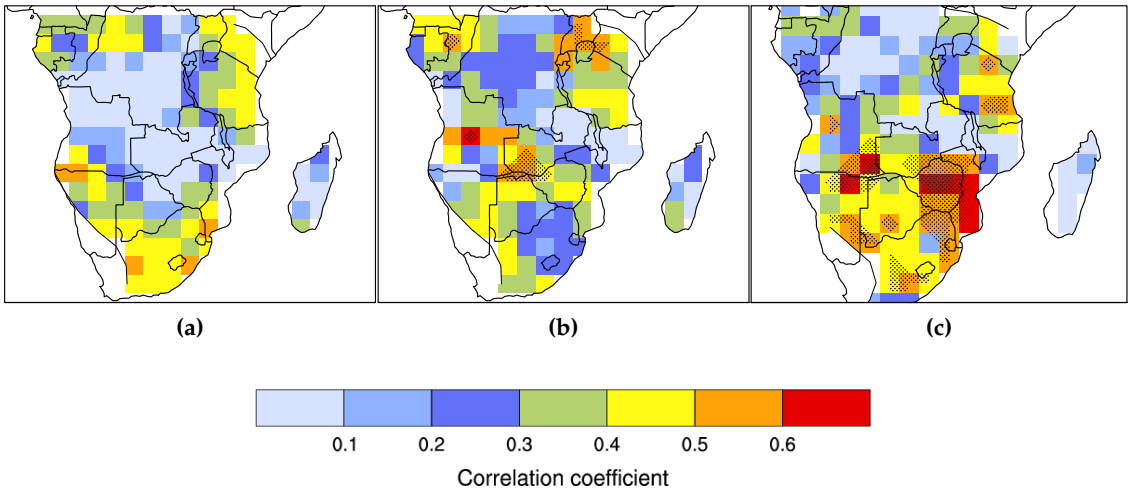


Figure 5.19: As figure 5.18, for ensemble mean precipitation vs GPCP. Only points are shown where the observed seasonal climatology is greater than 1mm/day.

almost-identical warm bias of two degrees in DEMETER and ENSEMBLES, whilst System 4 has on average a one degree cold bias for the island.

Precipitation biases are shown in figure 5.17. As for temperature, DEMETER and ENSEMBLES show a very similar pattern of bias, and again there is a large improvement with System 4. Both DEMETER and ENSEMBLES have a large wet bias of over four mm/day centralised over the south east, in the region of maximum rainfall for the season. Outside here there is a dry bias, particularly directly to the north east,

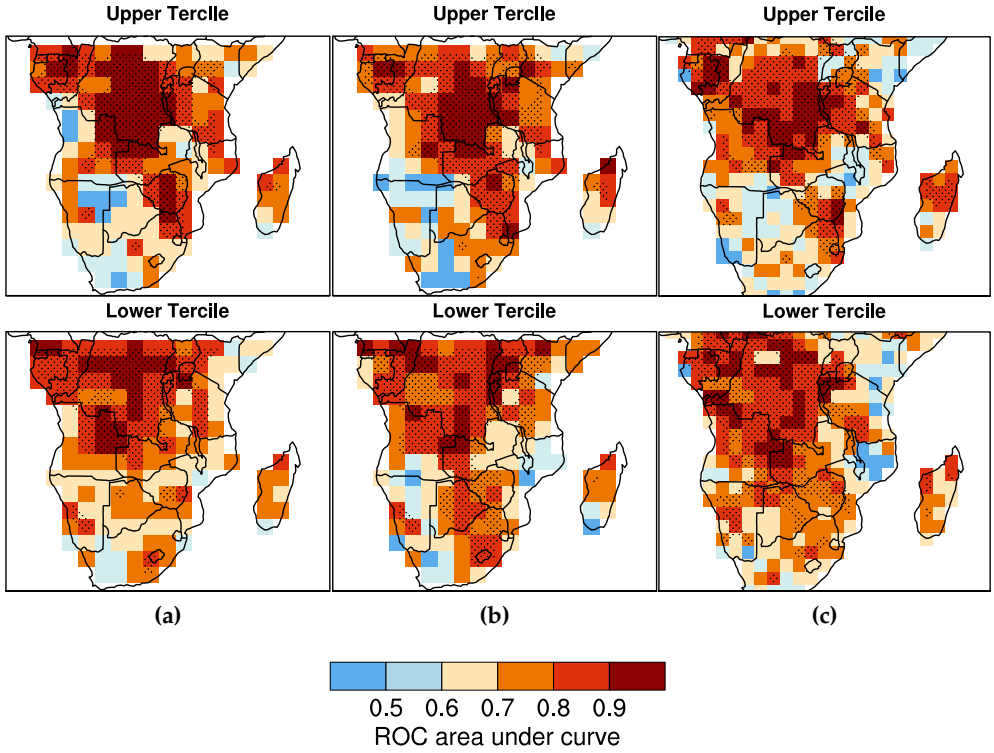


Figure 5.20: Relative operating characteristic area under curve (ROC AUC) for DJF temperature vs NCEP, for the November start dates of DEMETER, ENSEMBLES and System 4 (a-c). Stippled area indicates where the AUC is significant at the 95% level.

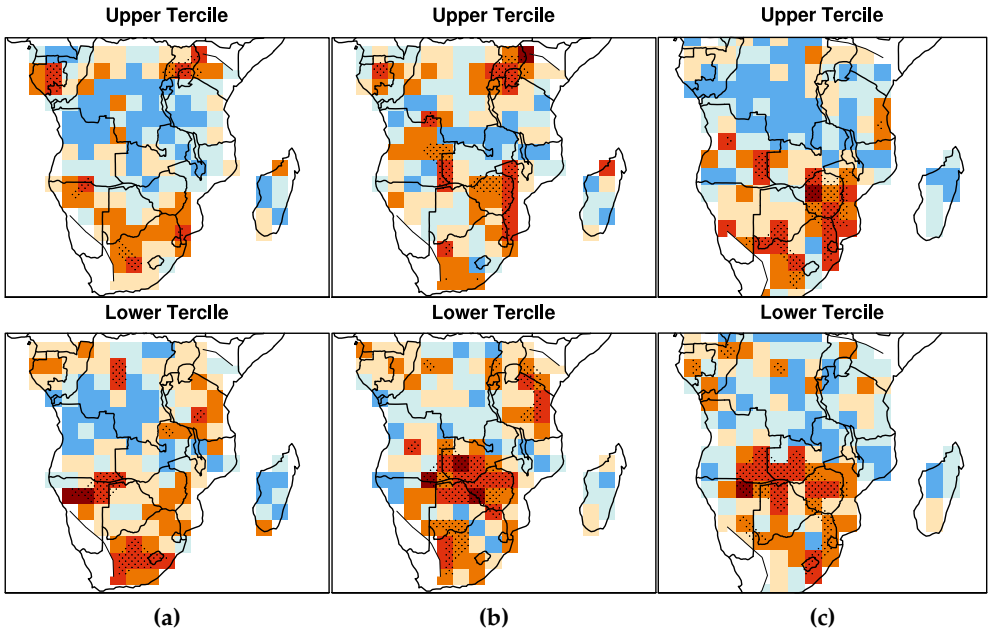


Figure 5.21: As figure 5.20, for precipitation vs GPCP. Only points are shown where the observed seasonal climatology is greater than 1mm/day.

which suggests that the rainfall is displaced. System 4 (figure 5.17c) has a much reduced bias; most of the region is only different from observations by one mm/day. For all three systems Madagascar has a large dry bias.

Correlations of ensemble mean temperature are shown in figure 5.18. The areas of significant correlation for DEMETER and ENSEMBLES cover most of the region, with a slightly larger area of significance for ENSEMBLES. Significant correlation for System 4 covers mostly the same region. The south western deserts, the area around lake Malawi and the north east are the only areas with correlation below significance.

For precipitation (figure 5.19) there is a large improvement between DEMETER and System 4. DEMETER has no significant correlation anywhere, with the largest values of correlation over South Africa. ENSEMBLES correlations are slightly higher than DEMETER and extend further inland, however only a few isolated gridpoints have correlation above significance. For System 4 correlations are highest, with a large coherent patch of significant correlation over the south east, near to the area of maximum rainfall in the season.

ROC AUC for temperature are shown in figure 5.20. The pattern is similar between the systems and generally follows that of the correlations seen in figure 5.18. The highest score generally occurs over the centre of the continent, over the Congo. Outside this region there is also an area of significant correlation over the south east, over Zimbabwe. The score is generally highest in DEMETER for upper tercile events, reducing slightly for ENSEMBLES and System 4.

For precipitation (figure 5.21), the ROC AUC is mostly below significance for most of the region in DEMETER, with a coherent patch of skill over the south east, in the same region where there is significant correlation. For ENSEMBLES however, there is also significant skill over a region covering Zimbabwe, Botswana and Western Zambia, mostly for lower tercile events; there is also a patch of significance for upper tercile events over Uganda. System 4 shows a similar pattern of to ENSEMBLES, though without any skill over Uganda.

Reliability plots for Botswana are shown in figure 5.22. For upper and for lower tercile forecasts, only System 4 has a positive BSS. It is not high however, and high probability forecasts are very overconfident. That is, for the cases when model forecast probabilities are 80% and above, the observed frequency of events is much lower. For precipitation (figure 5.23), again DEMETER and ENSEMBLES have no reliability and a BSS of less than zero. System 4 has a positive BSS, with all points on the reliability curve except for one laying inside the consistency bars.

Lack of skill here for Botswana is reflected in the value plots (figure 5.24 for

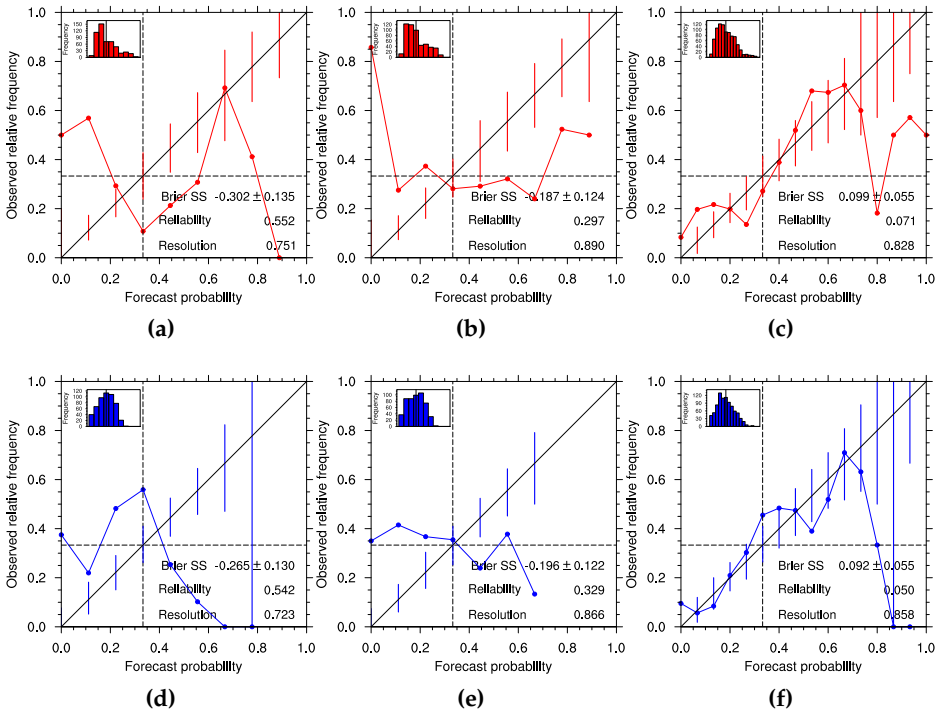


Figure 5.22: Reliability of upper (a-c) and lower (d-f) tercile DJF temperature forecasts over Botswana (vs NCEP): DEMETER (a & d), ENSEMBLES (b & e) and System 4 (c & f). Forecasts issued at the start of November.

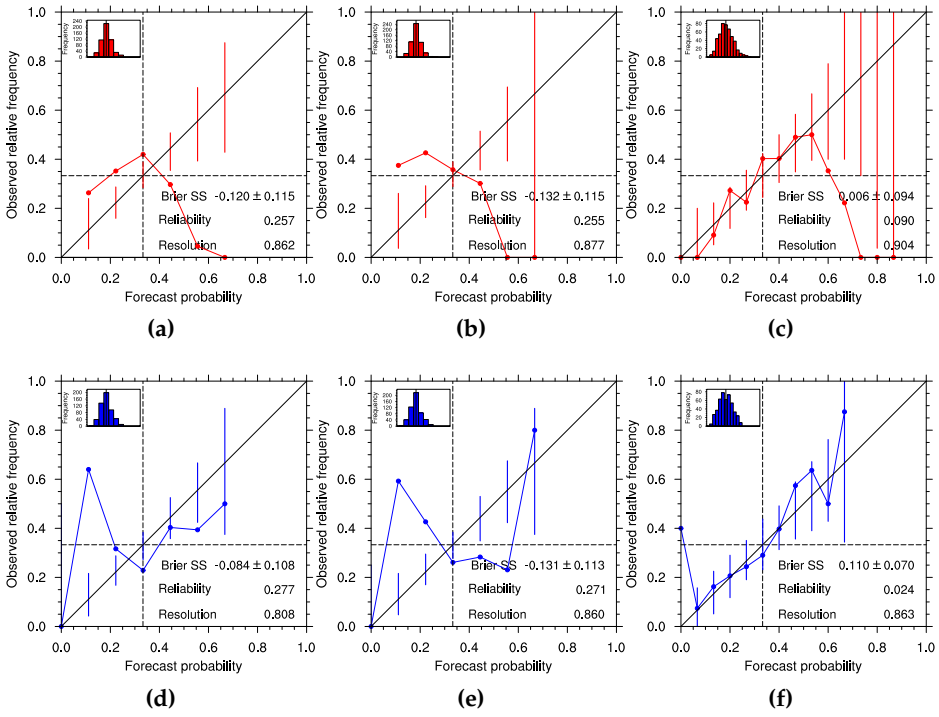


Figure 5.23: Reliability of Botswana precipitation vs GPCP, details as in figure 5.22.

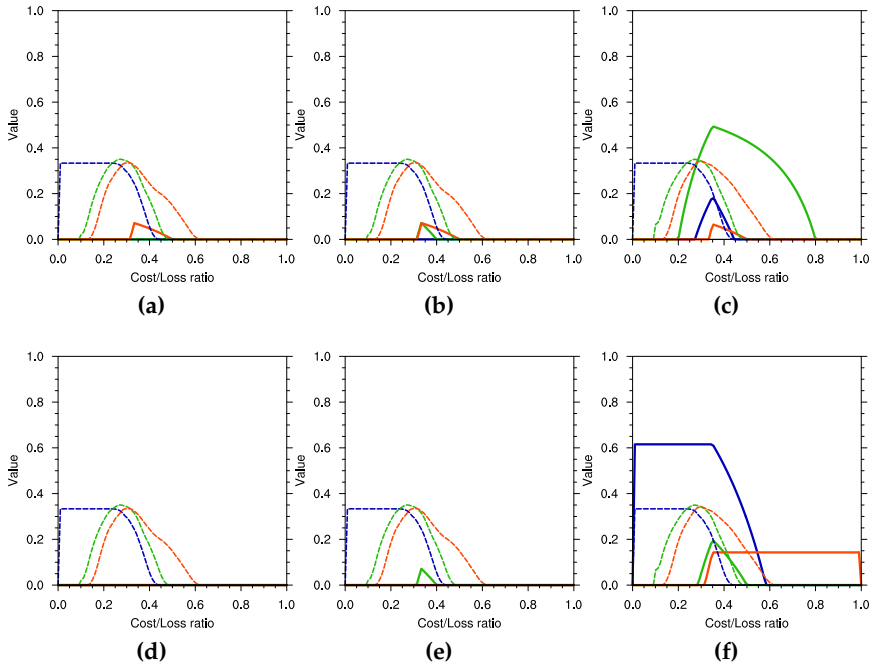


Figure 5.24: Value of upper (a-c) and lower (d-f) tercile JAS temperature forecasts over Botswana, for DEMETER (a & d), ENSEMBLES (b & e) and System 4 (c & f). Curves for 30%, 50% and 70% decision thresholds are shown by blue, green and orange lines with dashed lines indicating 95% significance level for each threshold. Forecasts issued at the start of November.

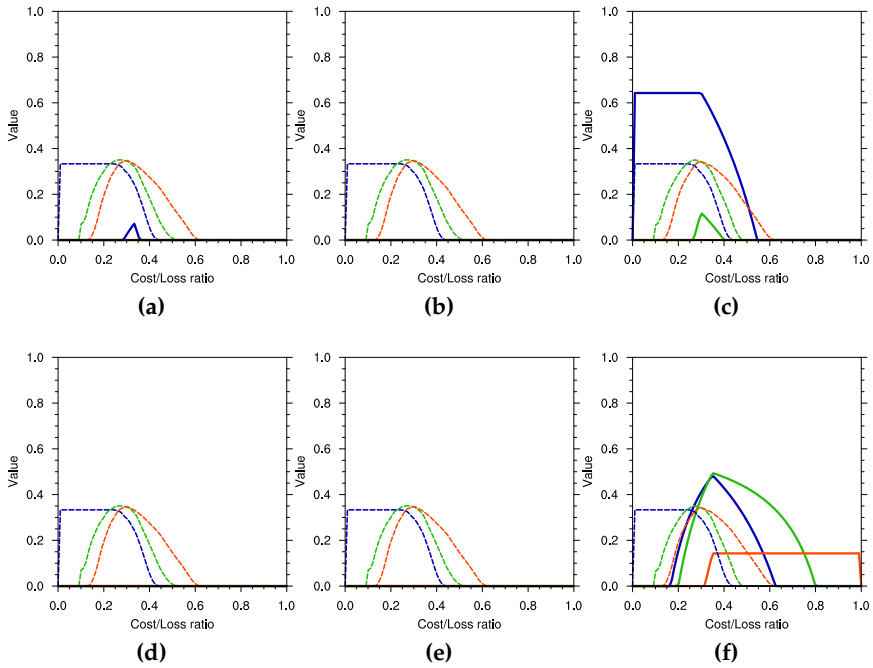


Figure 5.25: Value of Botswana precipitation vs GPCP, details as in figure 5.24.

temperature); for upper and for lower tercile events, DEMETER and ENSEMBLES show no value above significance. System 4 has value, with the 50% threshold curve above significance for upper tercile events, and both the 30 and the 70% curves above significance for lower tercile. For precipitation (figure 5.25) the results are similar, with DEMETER and ENSMEBLES showing no value above significance. System 4 again shows significant value for upper tercile precipitation at the 30% threshold, whilst for lower tercile all three thresholds give significant value.

Turning to the other southern Africa subregion, Malawi, reliability plots are shown in figure 5.26. BSS is highest for System 4, but including the error it is not different from zero. This is the case for upper and lower tercile forecasts, and is also true for both upper and lower tercile precipitation forecasts (figure 5.27).

Value plots for temperature and precipitation over Malawi are shown in figures 5.28 and 5.29. It can be seen that value is below significance for both upper or lower tercile events for both variables and all three systems. The only exception to this is System 4 for lower tercile precipitation (figure 5.29f), which has value just above significance for the 70% threshold.

5.2.3 An interpretation of biases

To conclude this section, it may be useful to consider what it is possible to learn from model validation, particularly biases. It is difficult however to pinpoint the reason for any specific model bias, but some general statements may be made.

A wet bias indicates too much rainfall: either rainfall events are too frequent, their magnitude is too high or a combination of both effects is occurring. Generally if there is too much rainfall too much moisture is being advected from nearby water bodies to the land, or the atmospheric conditions are conducive to moisture in the atmosphere falling as precipitation when in reality it may not. Too much advection of moisture suggests an excess of convection over nearby water, potentially due to either a warm sea surface temperature (SST) bias, or to problems in the model representation of convection.

One reason for a warm sea or land surface biases may be a too-strong introduction of water or air from a warmer region. For instance if the Gulf Stream were too strong in a climate model, the northern part of the north Atlantic ocean would have a warm SST bias. A too-warm surface may also be due to insufficient cloud cover during the day, preventing warming from shortwave radiation. Conversely, too much cloud cover during the night would preventing long-wave radiation from escaping from space and reduce cooling, causing a warm bias, that is, increasing the greenhouse effect.

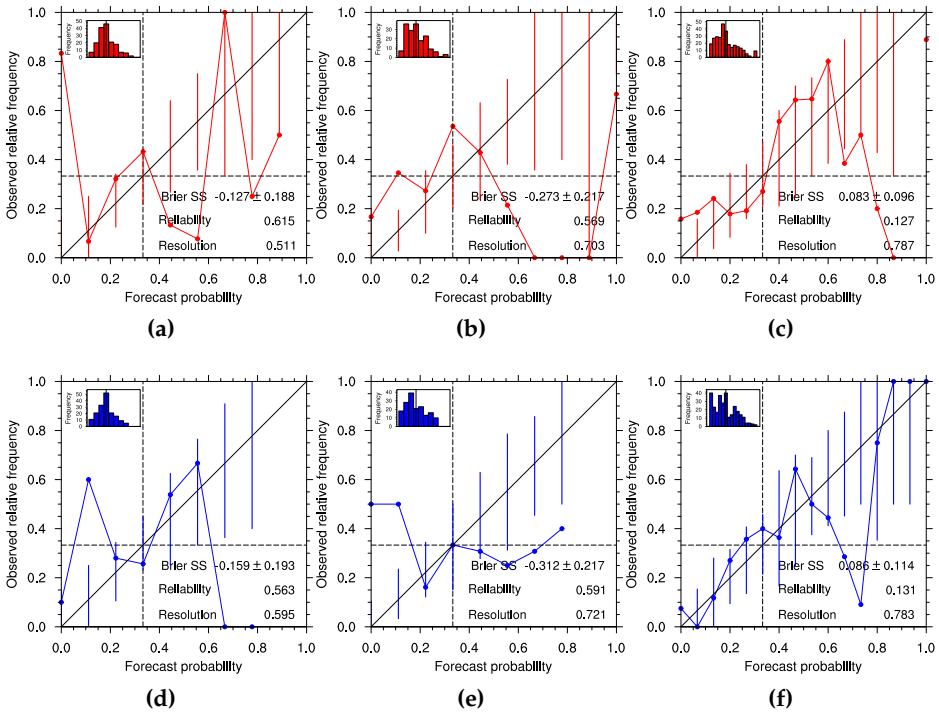


Figure 5.26: Reliability of upper (a-c) and lower (d-f) tercile DJF temperature forecasts over Malawi (vs NCEP): DEMETER (a & d), ENSEMBLES (b & e) and System 4 (c & f). Forecasts issued at the start of November.

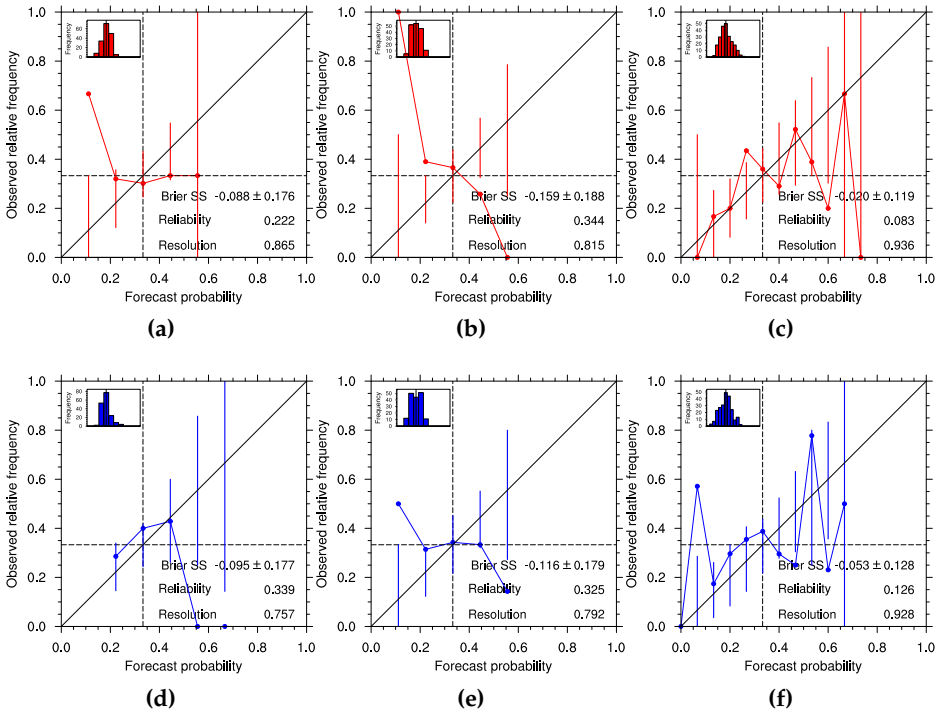


Figure 5.27: Reliability of Malawi precipitation vs GPCP, details as in figure 5.26.

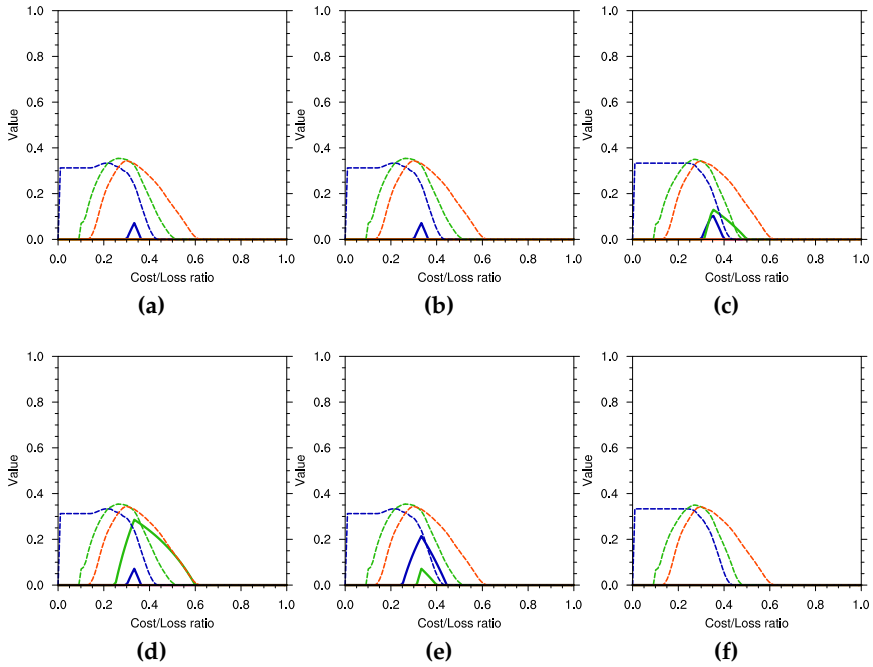


Figure 5.28: Value of upper (a-c) and lower (d-f) tercile JAS temperature forecasts over Malawi, for DEMETER (a & d), ENSEMBLES (b & e) and System 4 (c & f). Curves for 30%, 50% and 70% decision thresholds are shown by blue, green and orange lines with dashed lines indicating 95% significance level for each threshold. Forecasts issued at the start of November.

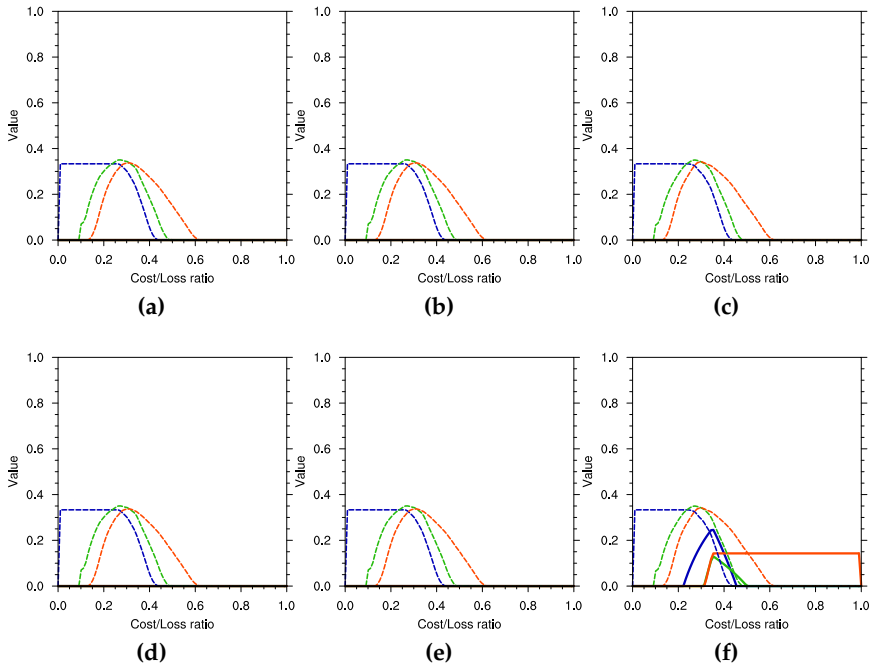


Figure 5.29: Value of Malawi precipitation vs GPCP, details as in figure 5.28.

There is also a link between temperature and precipitation biases through the cooling effect of precipitation. When there is too much precipitation on average during a season the surface cools, and an absence of precipitation would prevent cooling of the surface and cause a warm bias.

Biases will not necessarily link in this way: there are competing cooling effects from precipitation and warming effects from clouds at night (via the enhanced greenhouse effect mentioned above). Therefore enhanced (reduced) convection over a nearby water body would increase (decrease) cloud cover, potentially increasing (decreasing) surface temperature during the night and decreasing (increasing) surface temperature during the day via radiative effects; and if the conditions are right for precipitation, decreasing (increasing) the temperature via an enhanced (reduced) precipitation cooling effect. Finally, moisture at the surface allows more partitioning of the incoming radiation into latent heat, damping the heating. When moisture is absent (if there is a warm bias), more energy is available for sensible heating, and the surface will warm.

5.3 Discussion: the evolution of seasonal forecast skill

The DEMETER, ENSEMBLES and System 4 seasonal hindcasts have been compared in this chapter. For the regions shown here and in appendix B, it can be said that temperature and precipitation biases are generally lower in System 4, whilst the areas of significant correlations and ROC AUC for System 4 are generally largest.

In some places there is not much improvement between the systems (e.g. temperature ROC AUC over India, figure B.15), but for most regions the skill of ENSEMBLES is better than DEMETER, and the skill of System 4 is equal to and often higher than ENSEMBLES. This is important since DEMETER and ENSEMBLES were projects which created multi-model hindcasts as a one off experiment, whilst ECMWF regularly produces a forecast with System 4. That is, these results shows that it is possible to move from using the seasonal multi-model research systems to a single model system without significant reduction in forecast skill. These results may be of interest to decision makers and policy makers, for whom seasonal climate forecasts can inform effective climate change adaptation strategy (Washington et al., 2006).

Results for reliability depend on the target region, but for most areas System 4 forecast again perform the best. This is the case for BSS and for economic value; these results are summarised in table 5.5. Forecasts are generally poor (and no better than climatology) for the subregions selected over the Indian subcontinent, but for the Sahel and for Botswana System 4 can provide good forecasts[2]. For regions like the Sahel where many people live at risk from climate variability these skilful seasonal climate forecasts are potentially useful.

It is noteworthy that the state-of-the-art System 4 hindcasts generally give the best forecasts when compared against multi-model ensembles, despite coming from a single model and it is arguable that value of forecasts could be further improved by using a multi-model ensemble made up of System 4 and its contemporary models[3]. However whilst forecasts have improved since the early days of seasonal forecasting, these results show that model skill does not steadily increase for all regions with the evolution of climate models. It is then not the case that more money and development will necessarily lead to improvement of forecasts. There are some places for which forecasts may benefit more from climate model development and others where

[2]These results for System 4 are robust if the full hindcast period is used instead of the common period between the three systems (not shown). Furthermore, results for System 4 were compared with the single model from ECMWF used within DEMETER and ENSEMBLES, and almost unanimously System 4 shows improvement for all regions studied (also not shown).

[3]A combination of models in fact does run operationally at ECMWF. Known as EUROSIP, it uses the state-of-the-art versions of the seasonal models from the UK Met Office, ECWMF, Météo-France and NCEP (EUROSIP, 2013).

	Temperature				Precipitation			
	BSS		Value		BSS		Value	
Location	UT	LT	UT	LT	UT	LT	UT	LT
Sahel	E	S	S	S	S	ES	ES	-
Gulf of Guinea	ES	S	S	-	D	D	-	-
Malawi	S	S	-	-	-	-	-	-
Botswana	S	S	S	S	-	S	S	S
Kenya	S	E	-	DE	-	-	-	S
West India	DE	o	o	o	-	-	-	S
Bangladesh	DE	-	-	-	o	o	o	ES

Table 5.5: Summary of BSS and Value results. The system which has the highest BSS or value is identified with the letters D, E and S and colours yellow, blue and red for DEMETER, ENSEMBLES and System 4 respectively. When the BSS or value for two systems are equal and better than a third, combinations of letters and colors are used. White cells with 'o' indicate when all systems have roughly equal scores and are better than climatology, grey cells with '-' indicate when no system is better than climatology.).

prediction skill is and may remain low in the future, despite investment. Techniques exist which may improve forecasts, however the application of post-processing methods is outside the scope of this study[4].

There is are limits to prediction of seasonal climate (Palmer, 2000), though there is certainly room for improvement of models, and new pathways of research may improve seasonal climate models and their forecasts (e.g. Frenkel et al., 2012). It is unclear how close the current generation of seasonal climate models are to this limit and this is not an easy question to answer. Nevertheless, there has been a definite improvement in forecast skill since the first research into seasonal forecasting was undertaken, particularly in West Africa, a region where climate models have traditionally had problems (Ruti et al., 2011). Furthermore, the forecast systems creating this value have moved into a stage of routine simulation at forecasting centres rather than coming from stand-alone academic projects and so are potentially available to users.

An important point to highlight is that System 4 has forecasts initialized at the start of every month, compared to DEMETER and ENSEMBLES which have only four start dates per year. Since each System 4 forecast simulates the subsequent seven months, this means that forecasts for JAS can be potentially be issued from the beginning of March. These multiple lead times increase user flexibility with regard to acting on

[4]Model predictions may be improved by using post-processing methods such as model output statistics. This involves constructing a statistical model to relating well-simulated variables from dynamical models (for example sea surface temperature indices or 850hPa temperature) to variables of interest which are not simulated well by a dynamical model (Bouali et al., 2008).

forecast information and opens up forecasts to other users operating on multiple timescales. A forecast updating every month, with skill changing as the target is approached encourages an evolving, adaptable system of decision making. Cheaper preparation activities can be taken earlier with only uncertain information available, with more expensive protection measures only being carried out once a point is reached when forecasts are known to be more skilful. The following chapter explores this question of the variation of skill with lead time in System 4.

CHAPTER 6

When can useful forecasts be made? A closer look at ECMWF System 4.

ECMWF System 4 hindcasts are initialised once every month. Each forecast runs for seven months, allowing predictions to be made four months in advance of the start of a three month target. This section describes the skill of System 4 at these different lead times, along with a short final section on interpretation of economic value.

6.1 Methodology

In this chapter the hindcasts from ECMWF's seasonal prediction model (hereafter referred to as System 4) are explored in more detail. The methodology follows chapter 5, where System 4 was described, as were the target regions (these can be found in table 5.3). The datasets used for validation are again the NCEP reanalysis for temperature and GPCP for precipitation (see 3.1 for further details).

The work here looks at the System 4 forecasts made at four months ahead, two months ahead and at the start of the three month rainy season target. Again as in chapter 5, results for West and southern Africa are presented, whilst figures for East Africa and the Indian subcontinent are left to Appendix C.

6.2 Results

Results are summarised in table 6.1. Subsequently results for each region are described in separate sections.

6.2.1 West Africa

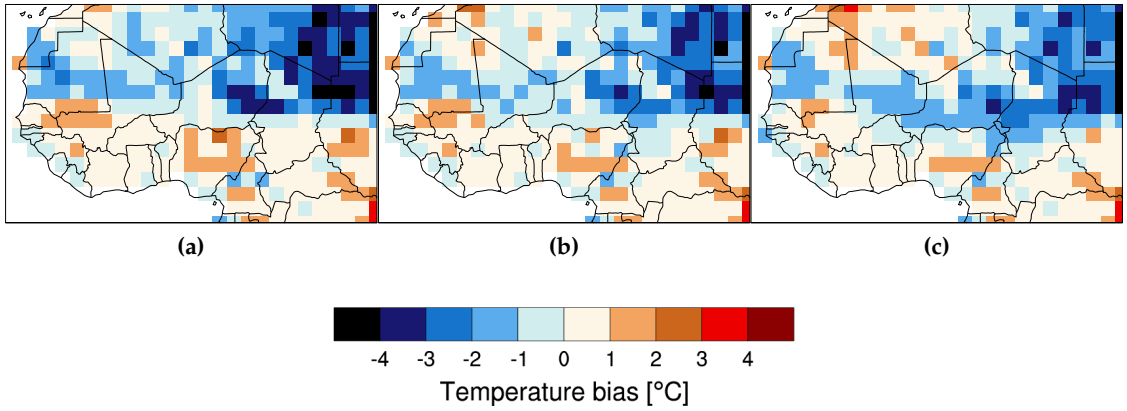


Figure 6.1: Ensemble mean JAS average temperature bias over West Africa vs NCEP, for System 4 forecasts issued March, May and July (a-c).

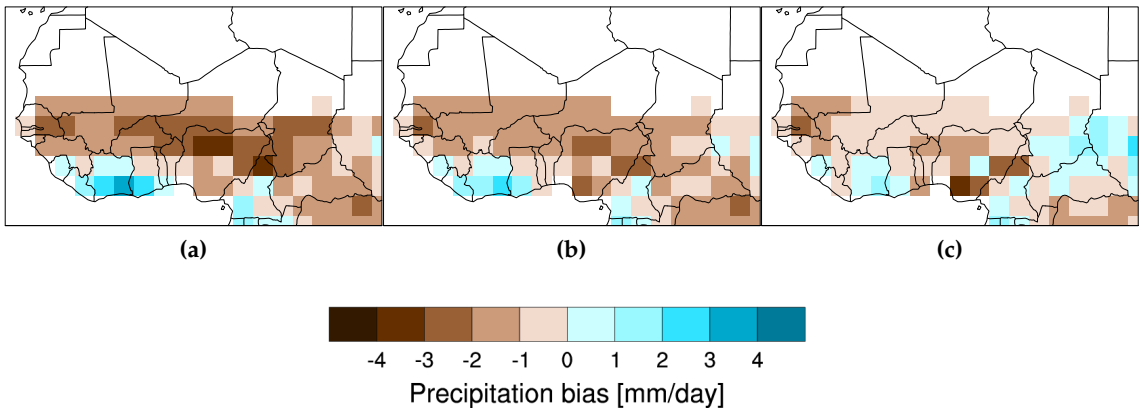


Figure 6.2: As figure 6.1, for ensemble mean precipitation vs GPCP. Only points are shown where the observed seasonal climatology is greater than 1mm/day.

The system 4 temperature bias over West Africa is shown in figure 6.1. The bias is warm but low near the coast, and cold over the desert in the north. At four months before the rainy season the bias is largest and as the target is approached it reduces. However the difference in bias between lead times is minimal.

	Temperature	Precipitation
West Africa		
Bias	Warm and low near the coast, little difference between lead times	Generally dry, reduction in bias at short lead time
Correlation	Significant at long lead times, increases as target approached	Below significance at long lead times, significant areas appear over Sahel and Ivory Coast as target approached
ROC AUC	Similar at all lead times, magnitude increases as target approached	Low at long lead times, significant at short lead times over Ivory Coast, Ghana and parts of Sahel
Southern Africa		
Bias	Coldest over desert; little difference between lead times	Slightly wet over most of the domain; little difference between lead times
Correlation	Increases with lead time; highest over Equatorial region and South Africa	Below significance at long lead time. At short lead very high over Tanzania
ROC AUC	Large area above significance, slight increase as target approached	Most of domain below significance at all lead times, at short lead times Tanzania, Botswana and South Africa have significant scores
East Africa		
Bias	No difference between start dates, complex pattern	Slightly dry, little difference between start dates
Correlation	Significant around the coast and north east Congo and north Tanzania. Small difference between start dates	Low at long lead times. Closest to the target significance over Uganda and Kenya
ROC AUC	Significant in the south of the region and around the coast; little difference between start dates	Below significance at all lead times, except for shortest lead forecasts, over part of Kenya
Indian Subcontinent		
Bias	Generally cold, slight improvement as target approached	Very dry over Bangladesh, low elsewhere. Little difference between start dates
Correlation	Above significance over west coast of India and Myanmar; magnitude increases with lead time	Below significance at all lead times, except for shortest lead, which is significant over tip of India and south of Nepal
ROC AUC	As for correlation: above significance over west coast of India and Myanmar; magnitude increases with lead time	As for correlation: below significance at all lead times, except for shortest lead, which is significant over tip of India and south of Nepal, particularly for upper tercile forecasts

Table 6.1: Summary of temperature and precipitation biases, ensemble mean correlation and ROC AUC for System 4 at different lead times. A summary of potential economic value can be found separately in table 6.2

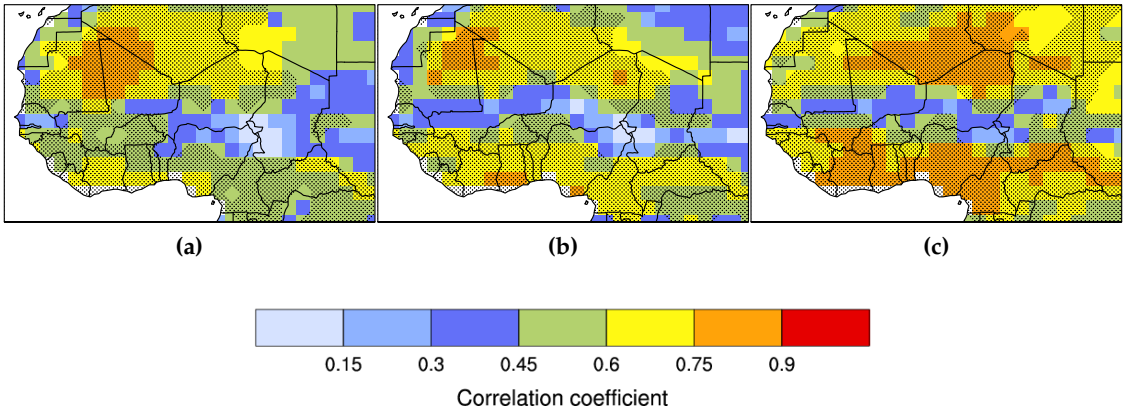


Figure 6.3: Pearson's product-moment correlations of JAS ensemble mean temperature vs NCEP, for System 4 forecasts issued March, May and July (a-c). Stippled area shows 99% confidence level, with sample size adjusted to take account for autocorrelation.

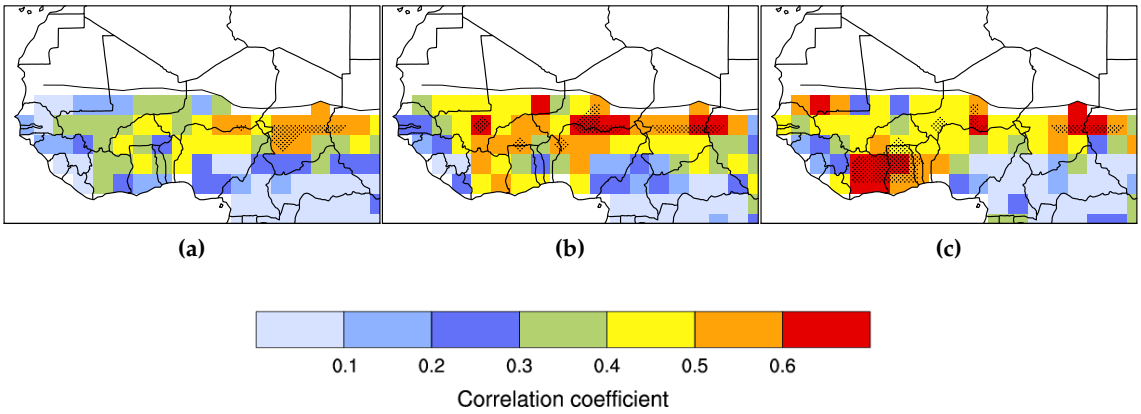


Figure 6.4: As figure 6.3, for ensemble mean precipitation vs GPCP. Only points are shown where the observed seasonal climatology is greater than 1mm/day.

Biases for precipitation are shown in figure 6.2. For most of the region System 4 is too dry, by one or two mm/day for the Sahelian region. A small region around coast of the Ivory Coast and Ghana is too wet by a similar amount. As the target approaches the magnitude of the bias reduces, so that forecasts made at the start of the rainy season have a bias of under one mm/day for most of the region.

Temperature correlations between the ensemble mean and reanalysis are shown in figure 6.3. The magnitude of the correlation coefficient increases as the target approaches, particularly near the coast. Forecasts issued at the start of the rainy season are significant everywhere, though a band over the Sahelian region is low for all start dates.

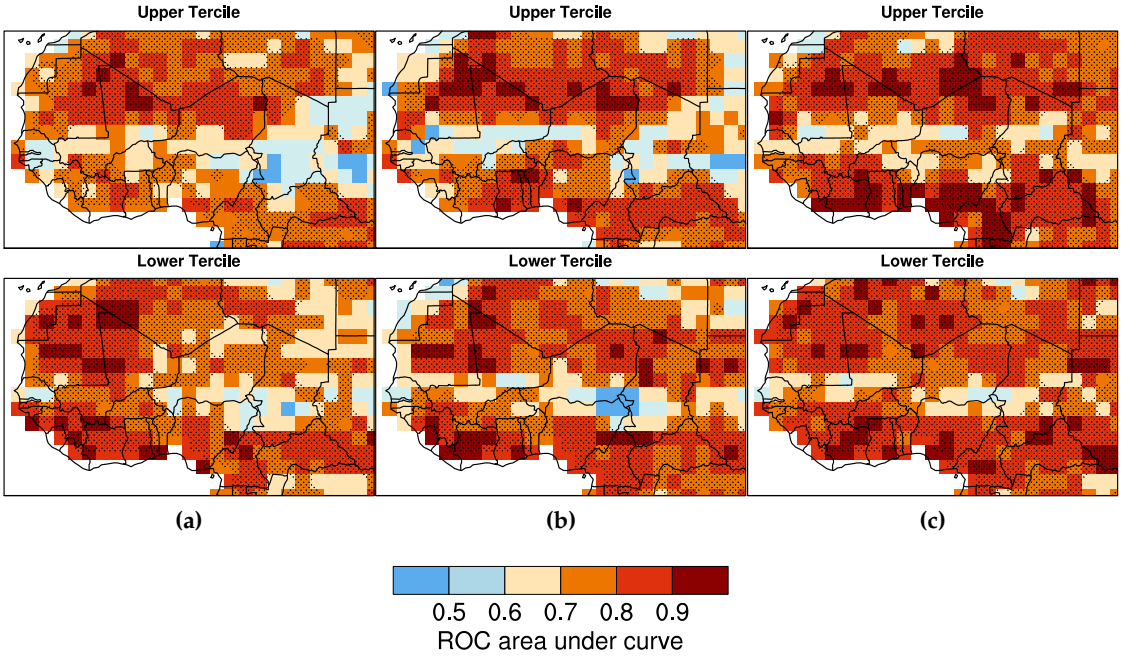


Figure 6.5: Relative operating characteristic area under curve (ROC AUC) for JAS temperature vs NCEP, for System 4 forecasts issued March, May and July (a-c). Stippled area indicates where the AUC is significant at the 95% level.

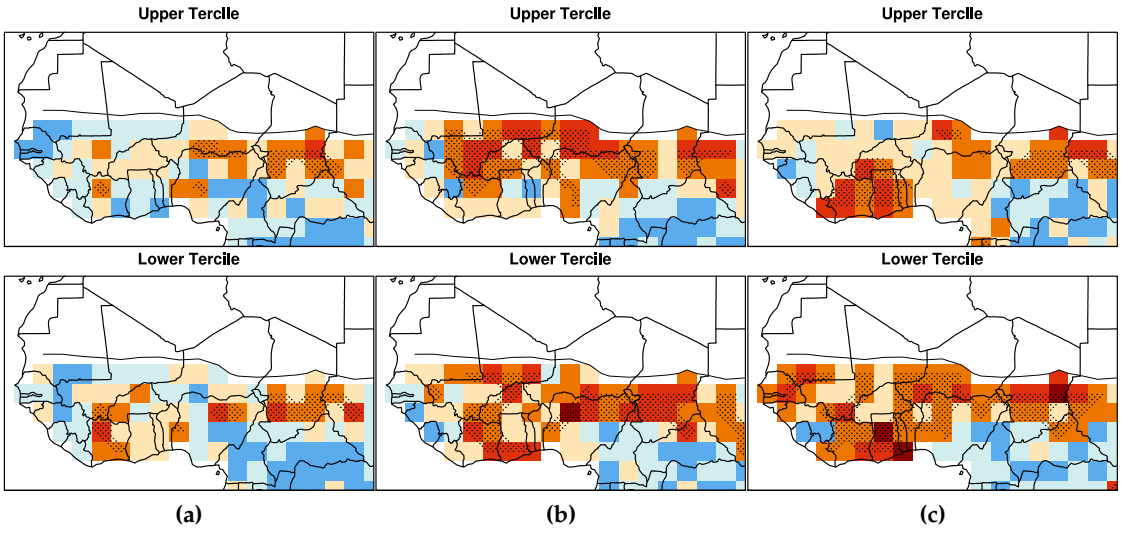


Figure 6.6: As figure 6.5, for precipitation vs GPCP. Only points are shown where the observed seasonal climatology is greater than 1mm/day.

For precipitation (figure 6.4) forecasts made furthest out from the rainy season have the lowest correlation, with only a few grid points displaying a correlation coefficient above significance. Two months out from the rainy season the correlation is larger, with increased correlation across the Sahel and larger number of significant grid points. Forecasts made at the start of the rainy season have a slightly reduced correlation over the Sahel, but do display a much larger correlation over the Ivory Coast.

ROC AUC for temperature is shown in figure 6.5. The spatial pattern is similar at all lead times, with high significance near the coast, reducing northwards, and high correlation over the desert. The magnitude of the score steadily increases as the target is approached; by the start of the rainy season forecasts have a ROC AUC over 0.9 for a large area over the coast. However there is still a band over the Sahel where the ROC AUC is relatively unchanged and below significance for all lead times; this is the same area which shows a low correlation for temperature in figure 6.3.

For precipitation (figure 6.6), ROC AUC is lowest for forecasts issued in March (figure 6.6a), with a few significant grid points near the coast. For forecasts initialised in May, the score is high over the Sahel for upper and lower tercile events. Upper tercile forecasts made at the start of the rainy season have a slight reduction in ROC AUC for the Sahel compared to forecasts initialised two months previously, whilst over the Ivory Coast and Ghana the score is much improved; a coherent area of skill is present in this region. For lower tercile forecasts at the start of the rainy season the correlation is above significance for a large area, over the Sahel and near the coast. There is perhaps an asymmetry between upper and lower tercile events due to the asymmetry in processes: for precipitation it is possible that the events leading to particularly high seasonal rainfall totals are not simply the opposite process for low totals; and the model may be able to simulate one set of processes more realistically better than the other.

Turning to the subregion defined as the Sahel in figure 5.1; reliability curves for JAS temperature are shown in figure 6.7. For upper tercile forecasts, the reliability curve is closest to the diagonal for forecasts at the longest lead times (for example, compare figures 6.7a to figure 6.7c), and the reliability reduces as the target is approached. For lower tercile forecasts this degradation in reliability is less, but for all lead times the forecasts are generally overconfident. Generally there is no change in the BSS as the target is approached. Without taking BSS errors into account the March forecasts have the highest score, though the difference in BSS between start dates is within the error bounds.

Reliability curves for Sahel precipitation are shown in figure 6.8. At the longest leadtime, the BSS is close to zero for upper and lower tercile events, with the reliability curve showing that there is no resolution for forecasts of lower tercile events. Forecasts at

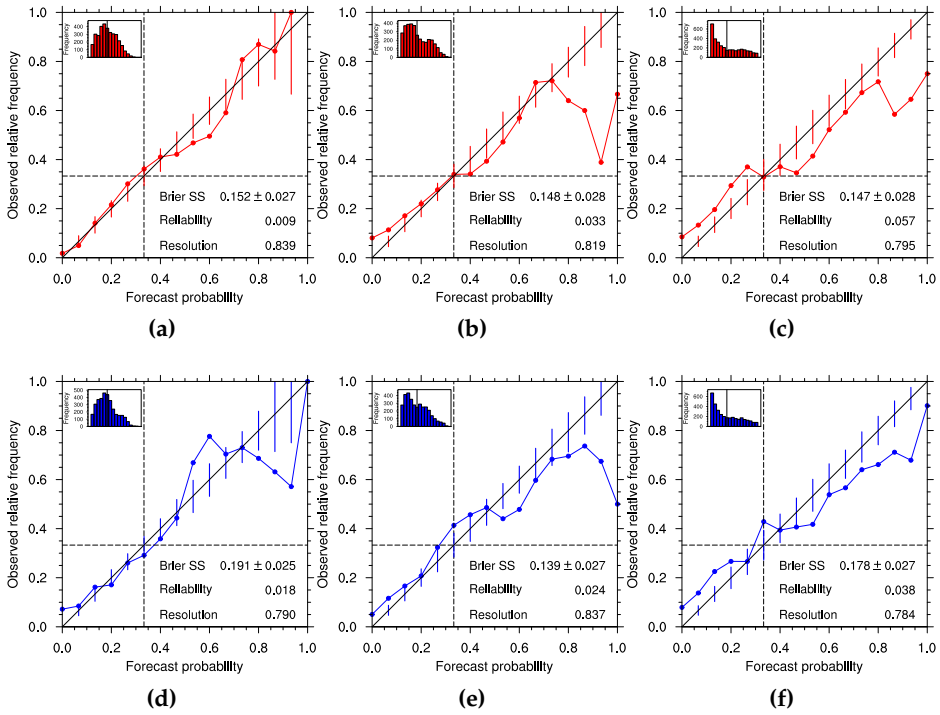


Figure 6.7: Reliability of upper (a-c) and lower (d-f) tercile JAS temperature forecasts over the Sahel region. Reliability is shown for System 4 forecasts issued March (a & d), May (b & e) and July (c & f).

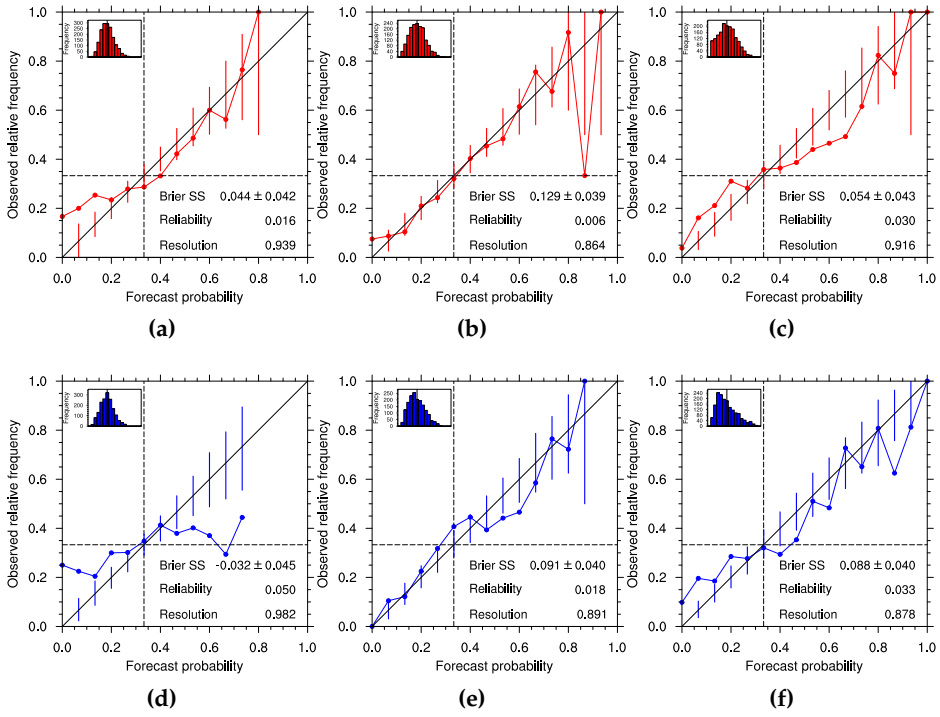


Figure 6.8: Reliability of Sahel precipitation vs GPCP, details as in figure 6.7.

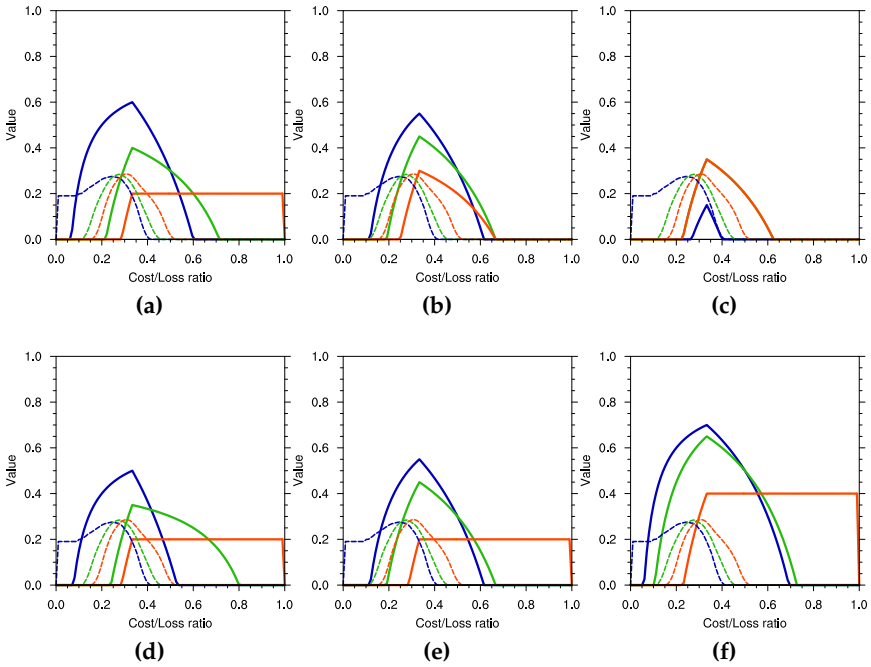


Figure 6.9: Value of upper (a-c) and lower (d-f) tercile JAS temperature forecasts over the Sahel, for System 4 forecasts issued March (a & d), May (b & e) and July (c & f). Curves for 30%, 50% and 70% decision thresholds are represented by blue, green and orange lines with dashed lines indicating 95% significance level for each threshold.

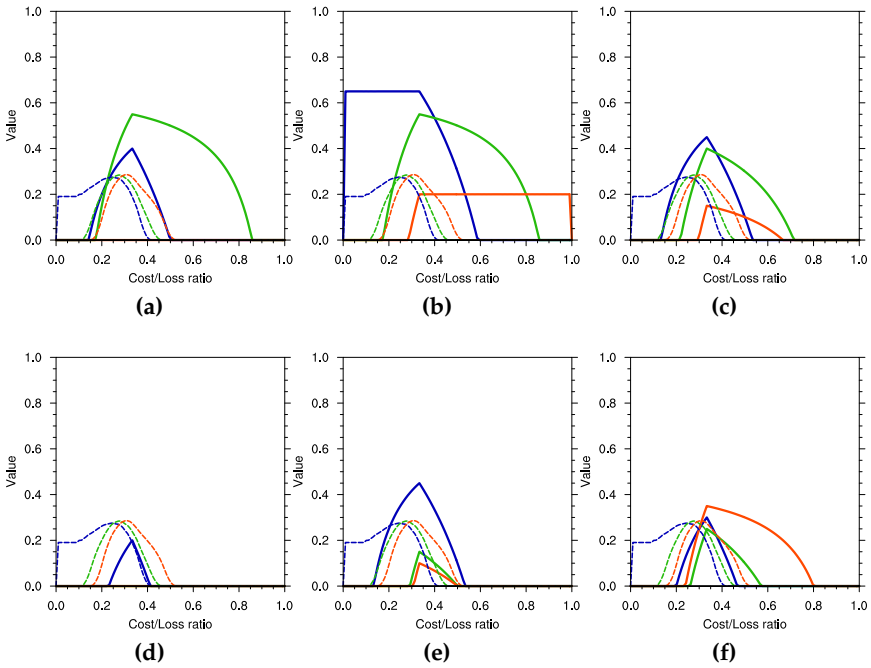


Figure 6.10: Value of Sahel precipitation vs GPCP, details as in figure 6.9.

two months before the rainy season are improved; showing the highest BSS for upper tercile events at all leadtimes. Comparing this with forecasts made at the start of the rainy season suggests that the reliability of the forecast reduces as the target is further approached; the reliability curve for the start of the rainy season (figure 6.8c) shows that forecasts are overconfident, with most points on the curve lying outside the consistency bars. For lower tercile forecasts, the reliability is not much reduced between two months and zero months ahead.

Looking at economic value over the Sahel, for temperature (figure 6.9), the curves suggest that forecasts for upper tercile temperature have value above significance for longer lead time forecasts, whilst forecasts at the start of the rainy season do not have value much above significance (figure 6.9c). For lower tercile events the value of the shortest lead forecasts is highest. Value is above significance at the longest lead time and is relatively similar for two month ahead forecasts, whilst the curves are higher for all decision thresholds for forecasts made at the start of the rainy season. The variation of value with lead time reflects that of the BSS score for temperature; March forecasts have the highest value and BSS for upper tercile forecasts, whilst lower tercile forecasts have the highest value and BSS in the July forecasts.

For precipitation (figure 6.10), the value of lower tercile forecasts is only slightly above significance for forecasts closest to the target, whilst those at the longest lead have no value above what might arise by chance. For upper tercile forecasts there is positive value at the longest lead time, continuing through to two month lead forecasts. The value of the forecasts closest to the target is lowest, and is only just above significance.

Reliability curves for the Gulf of Guinea are shown in figure 6.11. There is a steady increase in BSS for upper tercile forecasts as the target is approached, with the reliability curve for forecasts made in July lying completely within the consistency bars. For lower tercile forecasts there is no further improvement between May and July, with similar-looking reliability curves and BSS for the May and the July forecasts. For precipitation (figure 6.12) the reliability is much improved as the target is approached, with lower tercile forecasts improving between March and May but not significantly afterwards.

Economic value curves for temperature are shown in figure 6.13). There is a steady increase in value for upper tercile forecasts as the target is approached, with a high value for July forecasts (figure 6.13c). For lower tercile forecasts this is not the case, with July forecasts arguably offering the lowest value. For precipitation (figure 6.14), there is little value above significance in March, with a slight increase in value at May, mainly for upper tercile forecasts. July forecasts have the highest value, with upper tercile forecasts offering a large value across the cost/loss domain.

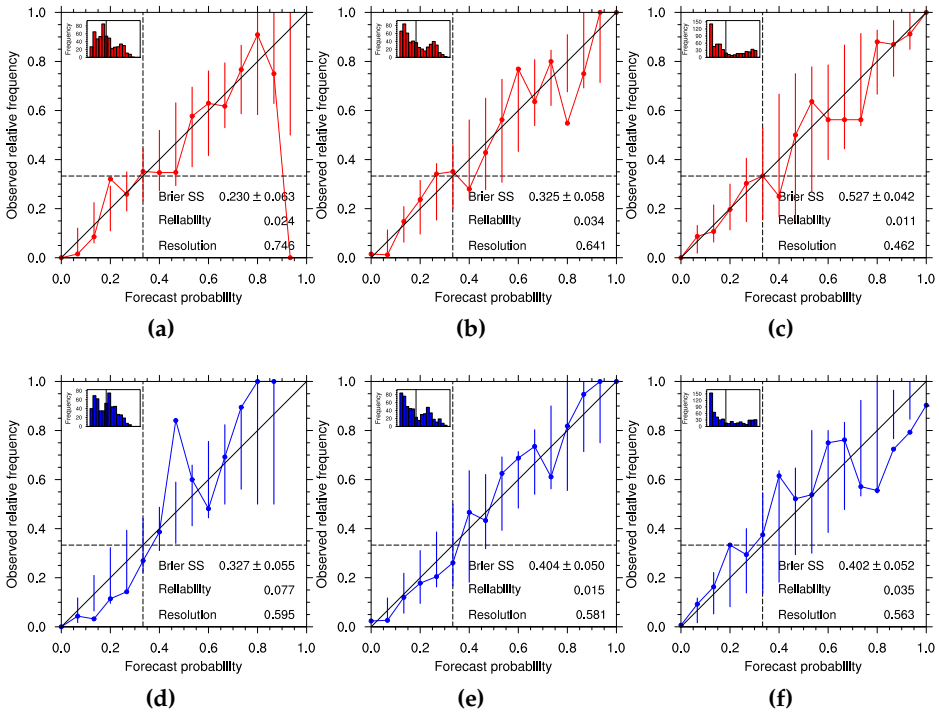


Figure 6.11: Reliability of upper (a-c) and lower (d-f) tercile JAS temperature forecasts over the Gulf of Guinea region. Reliability is shown for System 4 forecasts issued March (a & d), May (b & e) and July (c & f).

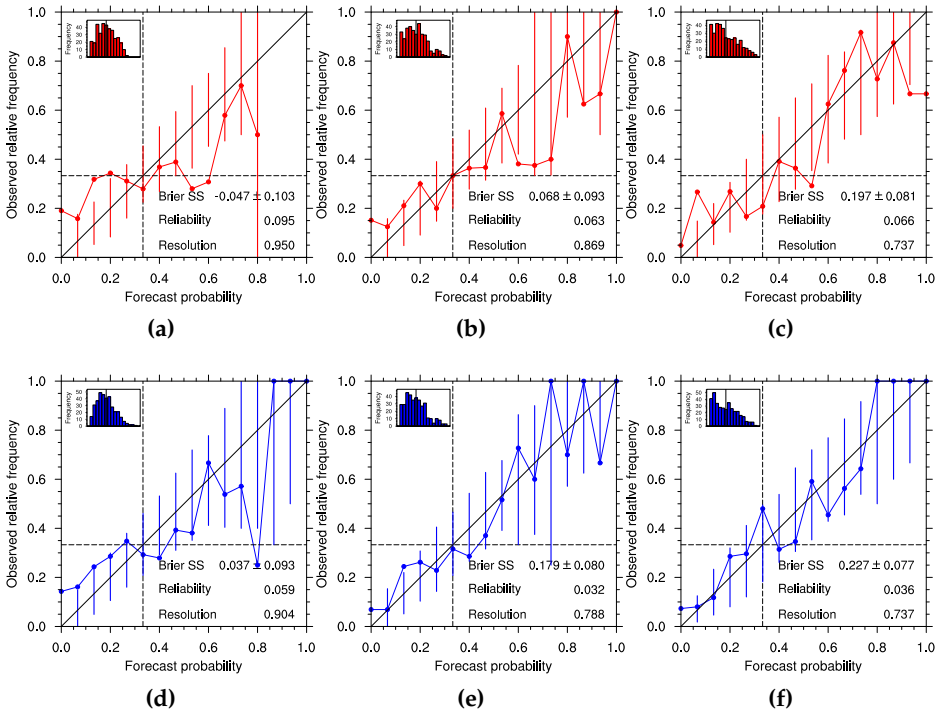


Figure 6.12: Reliability of Gulf of Guinea precipitation vs GPCP, details as in figure 6.11.

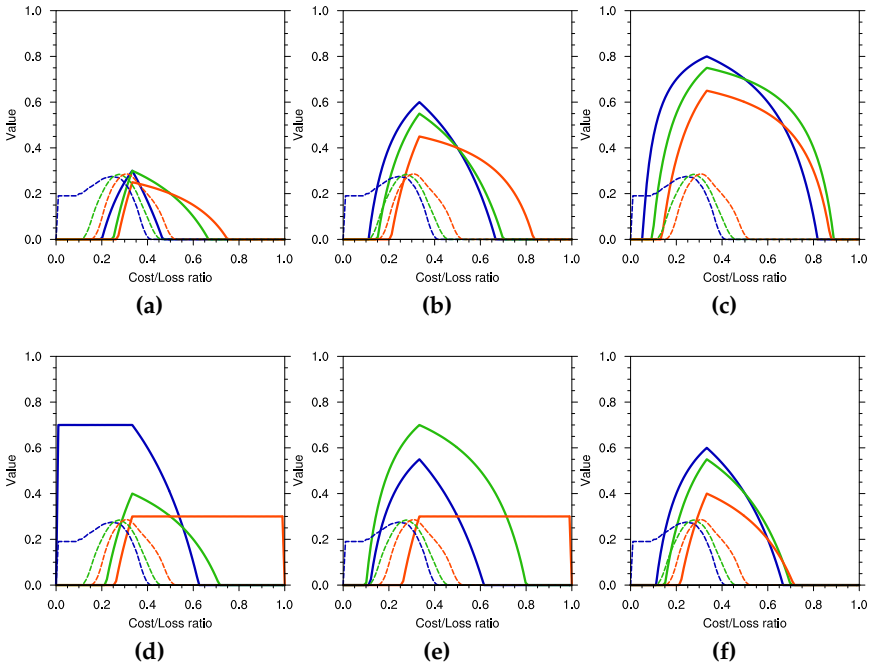


Figure 6.13: Value of upper (a-c) and lower (d-f) tercile JAS temperature forecasts over the Gulf of Guinea, for System 4 forecasts issued March (a & d), May (b & e) and July (c & f). Curves for 30%, 50% and 70% decision thresholds are represented by blue, green and orange lines with dashed lines indicating 95% significance level for each threshold.

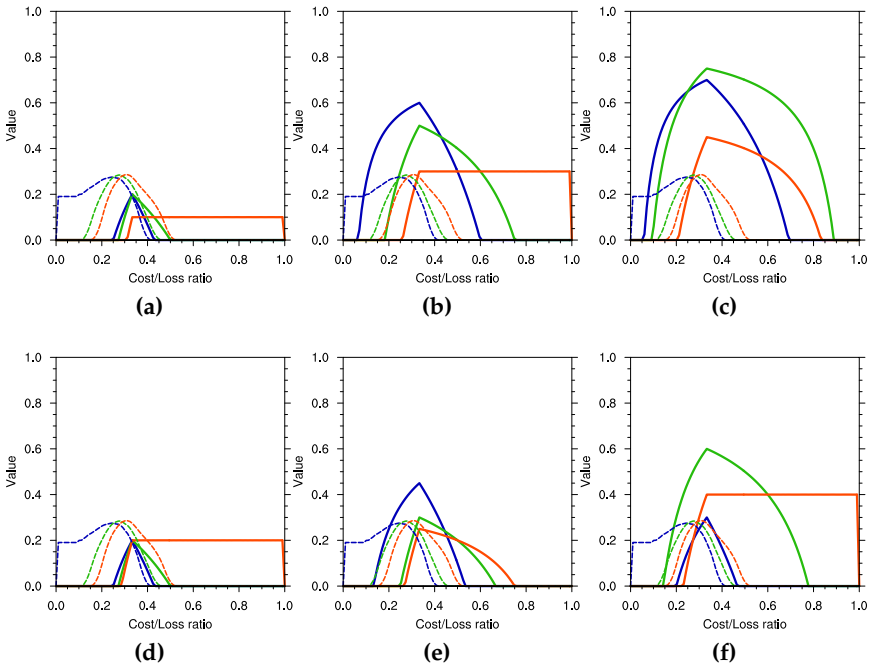


Figure 6.14: Value of Gulf of Guinea precipitation vs GPCP, details as in figure 6.13.

6.2.2 Southern Africa

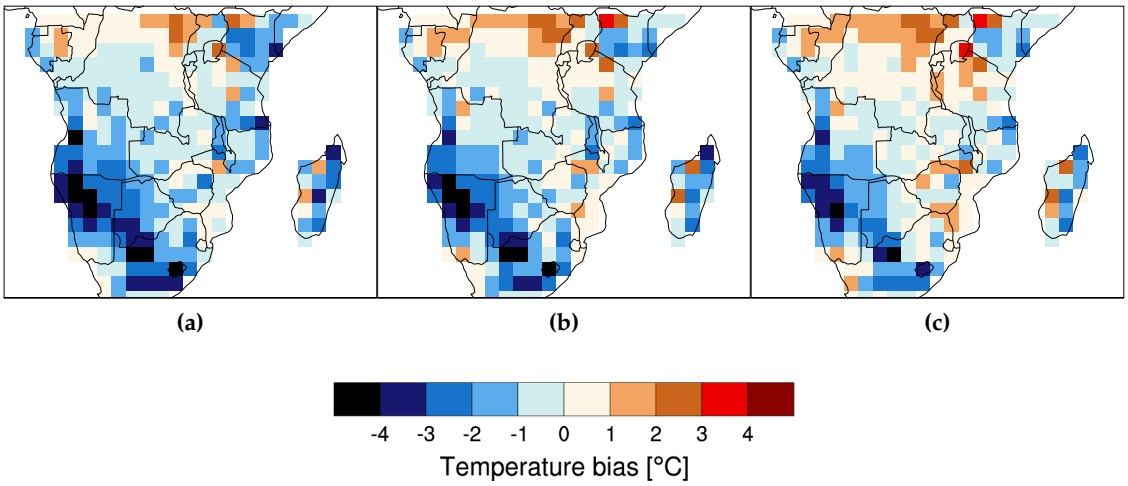


Figure 6.15: Ensemble mean DJF average temperature bias over southern Africa vs NCEP, for System 4 forecasts issued August, October and December (a-c).

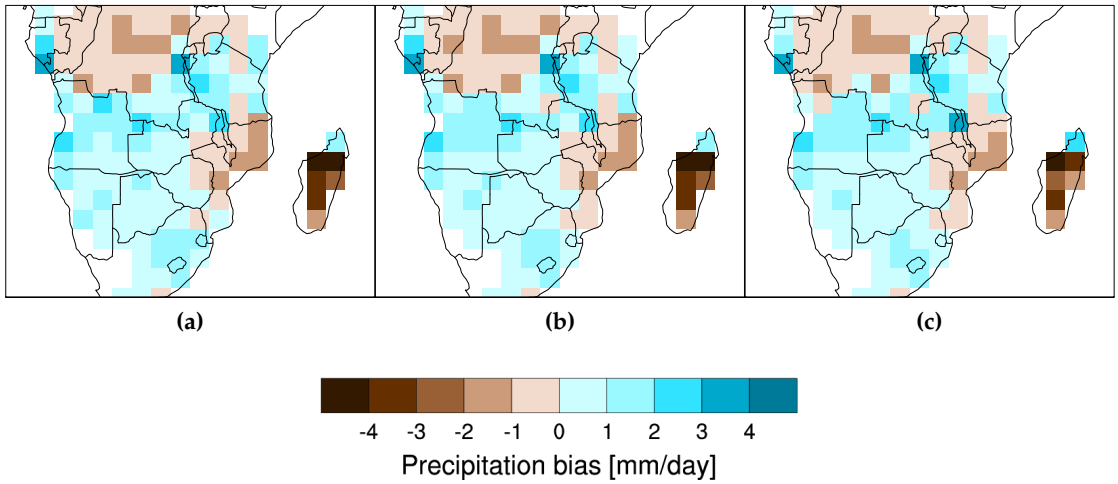


Figure 6.16: As figure 6.15, for ensemble mean precipitation vs GPCP. Only points are shown where the observed seasonal climatology is greater than 1mm/day.

Temperature biases for southern Africa DJF are shown in figure 6.15. System 4 is too cold over the south western deserts, with the bias reducing slightly as the target is approached. For the rest of the region the bias is similar for all start dates. The bias here is generally under one degree, except for Madagascar which has an average cold bias of around two degrees.

For precipitation (figure 6.16) the bias of forecasts issued in December is almost identical to forecasts issued in August, with most of the region under one mm per day too wet,

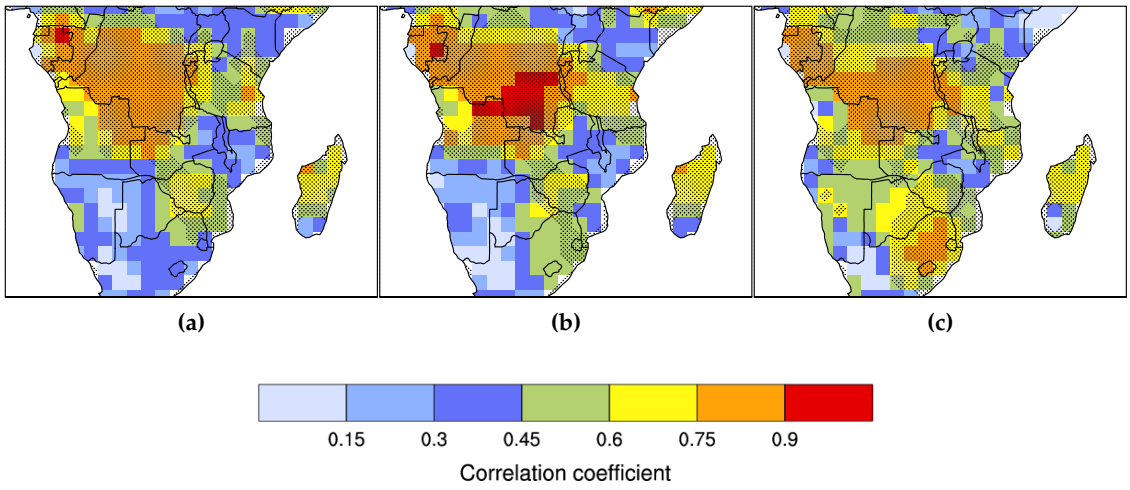


Figure 6.17: Pearson's product-moment correlations of DJF ensemble mean temperature vs NCEP, for System 4 forecasts issued August, October and December (a-c). Stippled area shows 99% confidence level, with sample size adjusted to take account for autocorrelation.

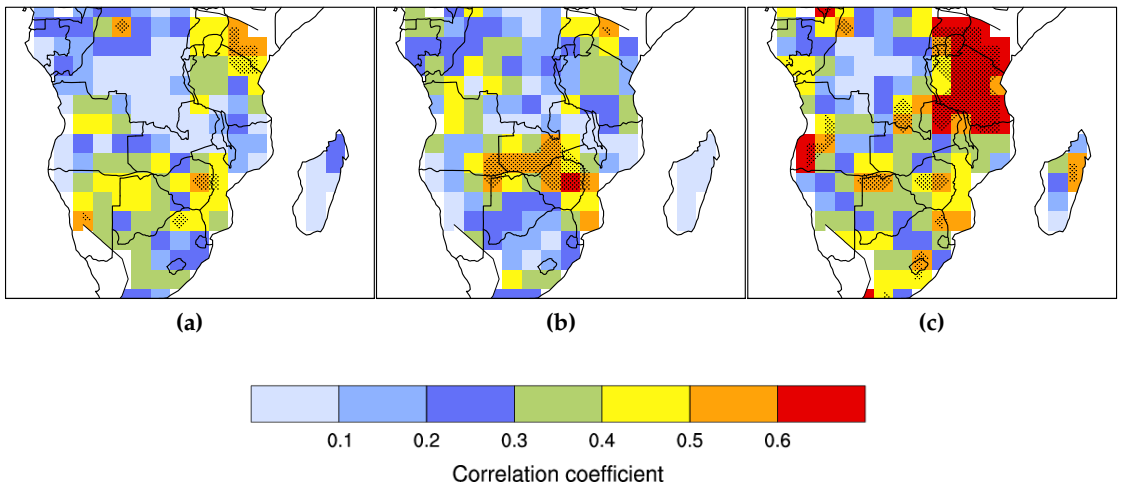


Figure 6.18: As figure 6.17, for ensemble mean precipitation vs GPCP. Only points are shown where the observed seasonal climatology is greater than 1mm/day.

whilst the model is slightly dry over Tanzania. The largest bias is over Madagascar, which does not receive enough rainfall by roughly two to three mm per day.

Temperature correlations are shown in figure 6.17. The area of correlation coefficient above significance increases as the target is approached. The largest increase in correlations occurs in north east of South Africa, where correlations increase from around 0.3 in August to over 0.75 in December. In central Africa, over the Congo, correlations are high in August, increase by October but reduce for forecasts made in December. For most of the region however, correlations are largest for forecasts made

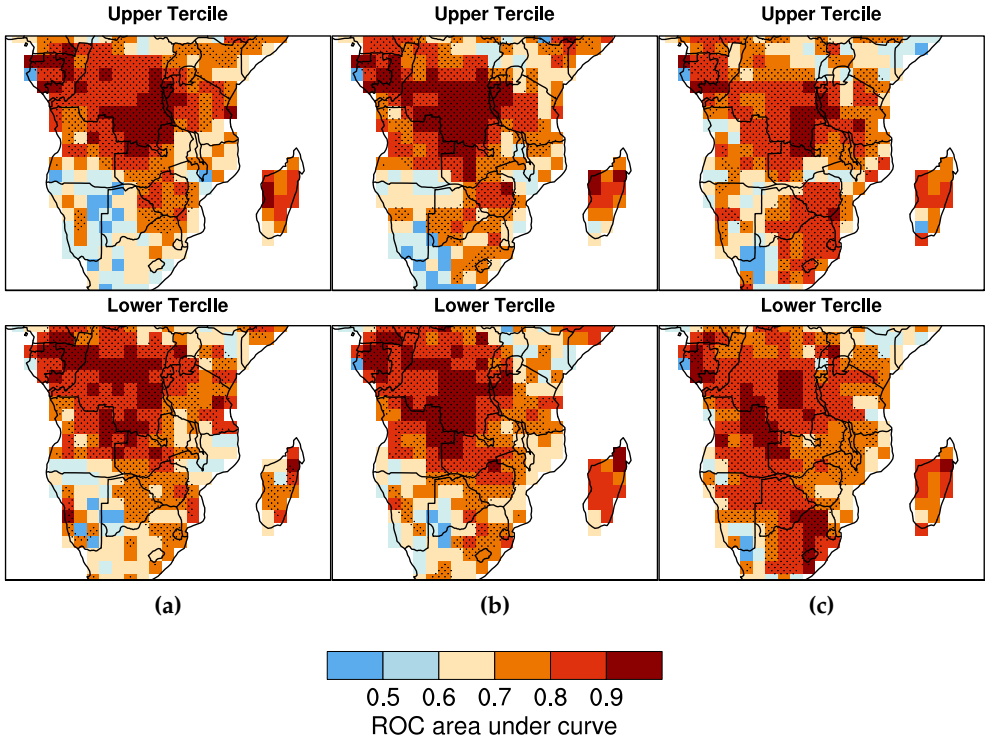


Figure 6.19: Relative operating characteristic area under curve (ROC AUC) for DJF temperature vs NCEP, for System 4 forecasts issued August, October and December (a-c). Stippled area indicates where the AUC is significant at the 95% level.

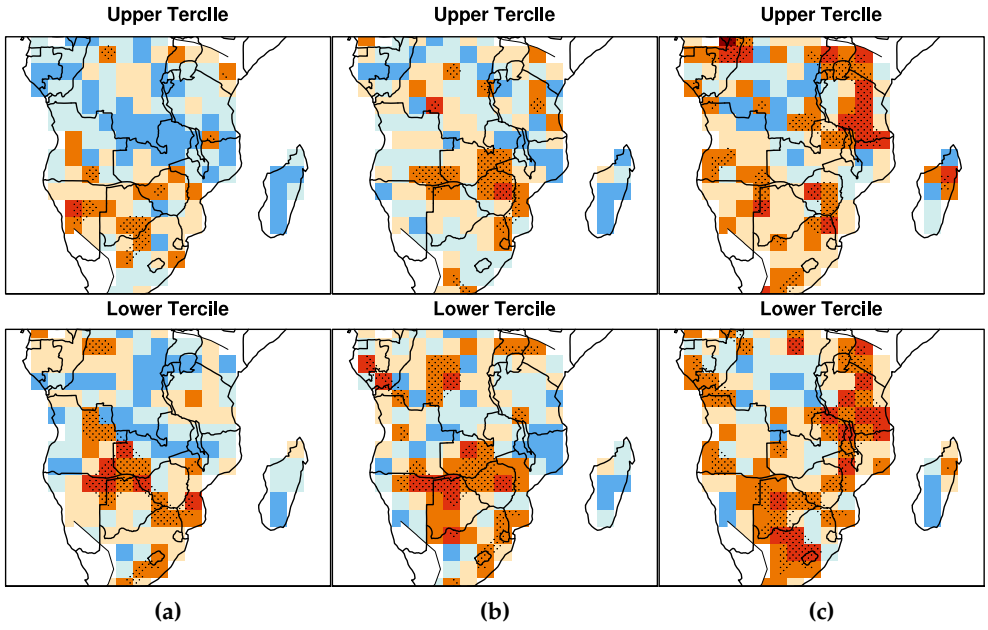


Figure 6.20: As figure 6.19, for precipitation vs GPCP. Only points are shown where the observed seasonal climatology is greater than 1mm/day.

closest to the target.

For precipitation (figure 6.18), correlations increase sharply in the region surrounding Tanzania between October and December forecasts, from 0.3 to over 0.6. For the rest of the region the coefficient is similar at all lead times, except for a region around Zimbabwe where correlations are highest in October before dropping by December.

Maps of ROC AUC for temperature are shown in figure 6.19. The area of significant ROC AUC is largest in December, particularly so for lower tercile event forecasts. At earlier lead times the score is roughly similar, though there is a slight reduction in the magnitude of the score over the Congo as the target is approached, whilst over the north east of South Africa the score increases the most, from around 0.6 in August to over 0.8 in December.

Precipitation ROC AUC maps are shown in figure 6.20. There is a slight increase in the magnitude of the score across most of the region, though in December most grid points do not have scores above significance. However, there are patches of significant ROC AUC in December over some areas, notably Tanzania, and further south, over Botswana/South Africa for lower tercile forecasts. For August and October forecasts there is some significant skill for lower tercile events in a region around the meeting point of Zimbabwe, Botswana and Zambia. However in December this significant skill is no longer present.

Reliability plots for temperature over Botswana are shown in figure 6.21. The reliability increases steadily as the target is approached, as does the BSS, from 0.032 in August to 0.177 in December for upper tercile forecasts, and from 0.105 to 0.267 for lower tercile. There is also a noticeable shift in the distribution of forecast probabilities, at longer lead times the probabilities cluster around the climatological frequency and as the target is approached the number of lower probabilities forecasts increases. That is, there is more variability in the probabilities from the model. Whilst in August the majority of the forecast probabilities cluster around the baseline probability (i.e. 33%) with few instances of high and low forecast probability, by December there more spread; there are more instances when over 60% of ensemble members indicate event, as well as instances where probabilities are less than 20%.

For precipitation (figure 6.22) the BSS is only slightly above zero for upper tercile forecasts, whilst for lower tercile forecasts the score is higher. At high forecast probabilities the reliability is lower, though at lower probabilities most of the reliability curve lies inside the consistency bars.

Economic value curves of temperature forecasts over Botswana are shown in figure 6.23. For August, forecasts of upper tercile events have no value above significance,

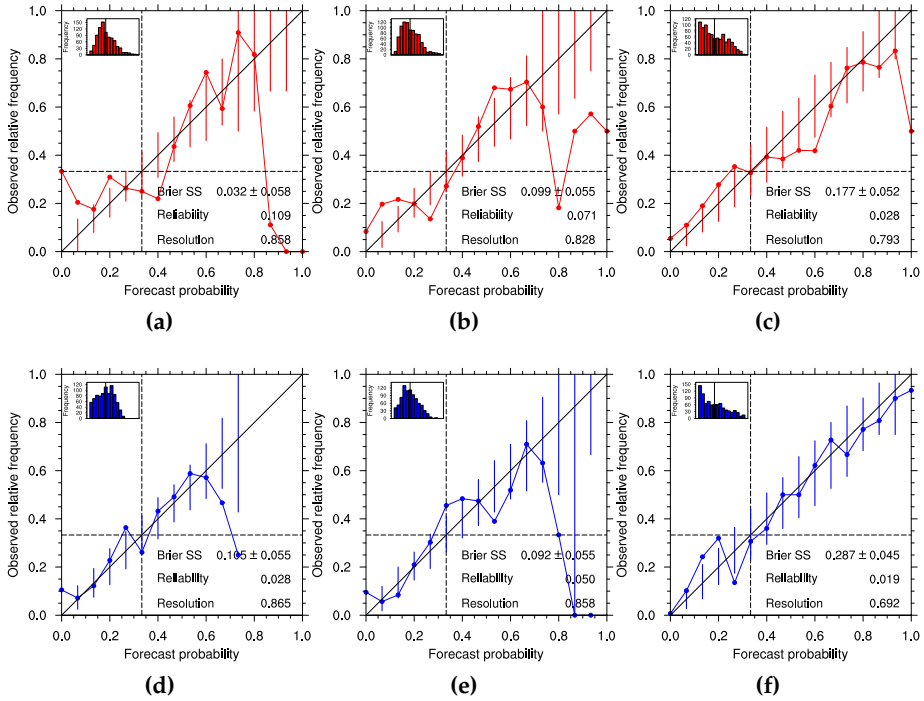


Figure 6.21: Reliability of upper (a-c) and lower (d-f) tercile DJF temperature forecasts over Botswana. Reliability is shown for System 4 forecasts issued August (a & d), October (b & e) and December (c & f).

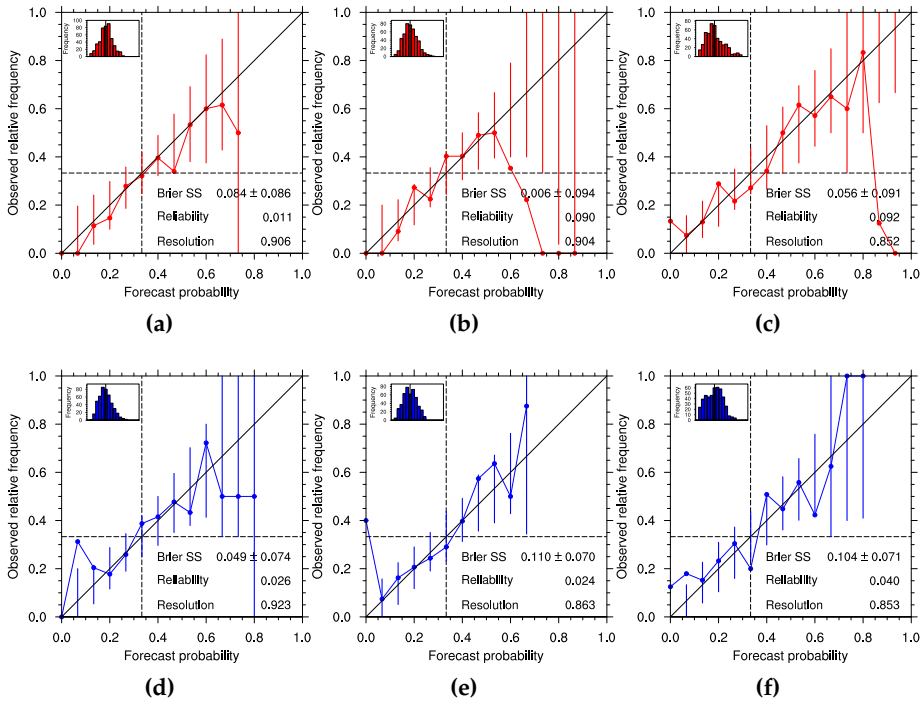


Figure 6.22: Reliability of Botswana precipitation vs GPCP, details as in figure 6.21.

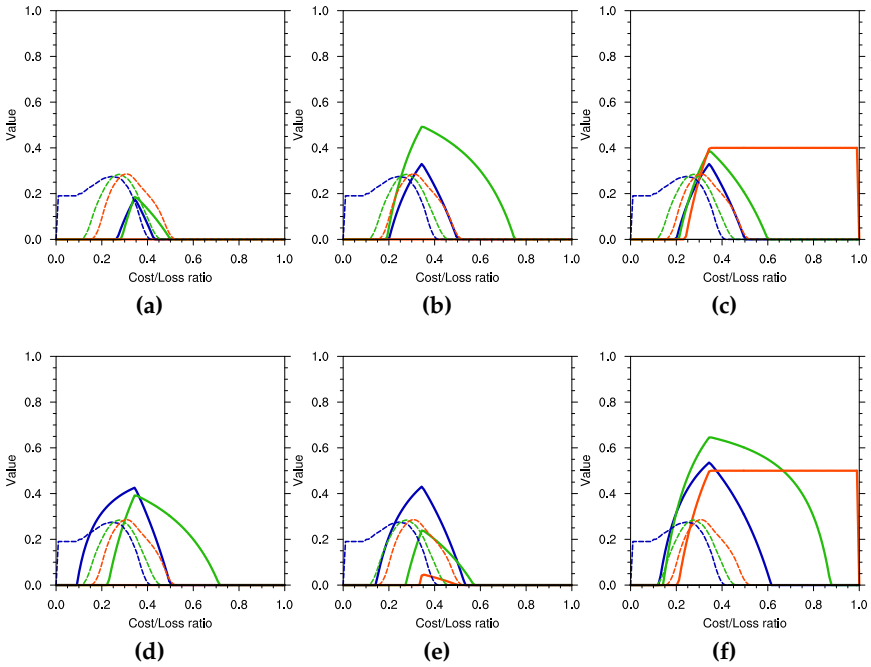


Figure 6.23: Value of upper (a-c) and lower (d-f) tercile JAS temperature forecasts over Botswana, for System 4 forecasts issued August (a & d), October (b & e) and December (c & f). Curves for 30%, 50% and 70% decision thresholds are represented by blue, green and orange lines with dashed lines indicating 95% significance level for each threshold.

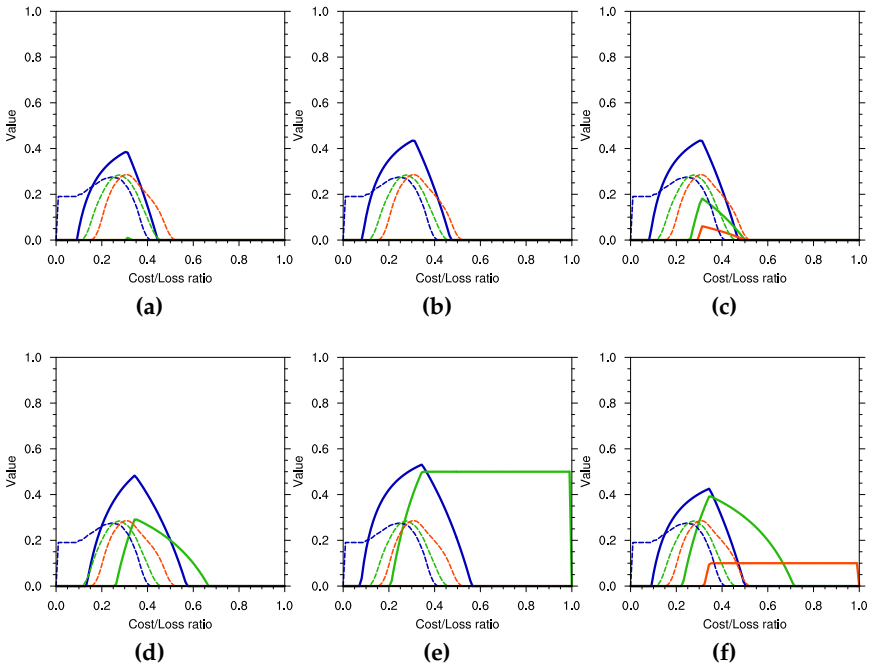


Figure 6.24: Value of Botswana precipitation vs GPCP, details as in figure 6.23.

whilst in October and December forecasts have some value above significance for some decision thresholds. For lower tercile events, August forecasts have some value above significance, which is roughly similar in October. By December the value increases significantly, with value above significance across the entire cost/loss domain for all decision thresholds. For precipitation (figure 6.24) the value of upper tercile forecasts stays roughly constant for all lead time forecasts; for the 30% threshold the value is slightly above significance. For lower tercile events, value is highest for August forecasts, whilst the value of December forecasts is reduced.

Reliability plots for temperature over Malawi are shown in figure 6.25. For upper tercile events the reliability is somewhat reduced between August and December, with BSS of close to zero. By December (figure 6.25c), upper tercile forecasts have a much improved reliability and BSS. The distribution of forecasts also changes significantly, with most points issuing a zero probability and the rest evenly distributed. For lower tercile forecasts this improvement in December is not observed, with the reliability of forecasts staying roughly constant. For precipitation (figure 6.26), the reliability curve for forecasts of upper and lower tercile events made in August and October lies far from the diagonal, though with most points still inside the consistency bars. The BSS is negative. For December forecasts the reliability is improved, with points lying closer to the diagonal, and the BSS positive. For December forecasts the BSS is largest for lower tercile events (6.25f).

Economic value curves for temperature are shown in figure 6.27. For August forecasts the value is only just above significance for upper tercile events and below for lower tercile. October forecasts have a slightly increases value, whilst December forecasts have a much larger value, particularly for lower tercile forecasts. For precipitation (figure 6.28), upper tercile forecasts have no value above significance at any lead time. Lower tercile event forecasts have little skill above significance for August and October start dates, whilst forecasts initialised in December have a large value for lower tercile forecasts.

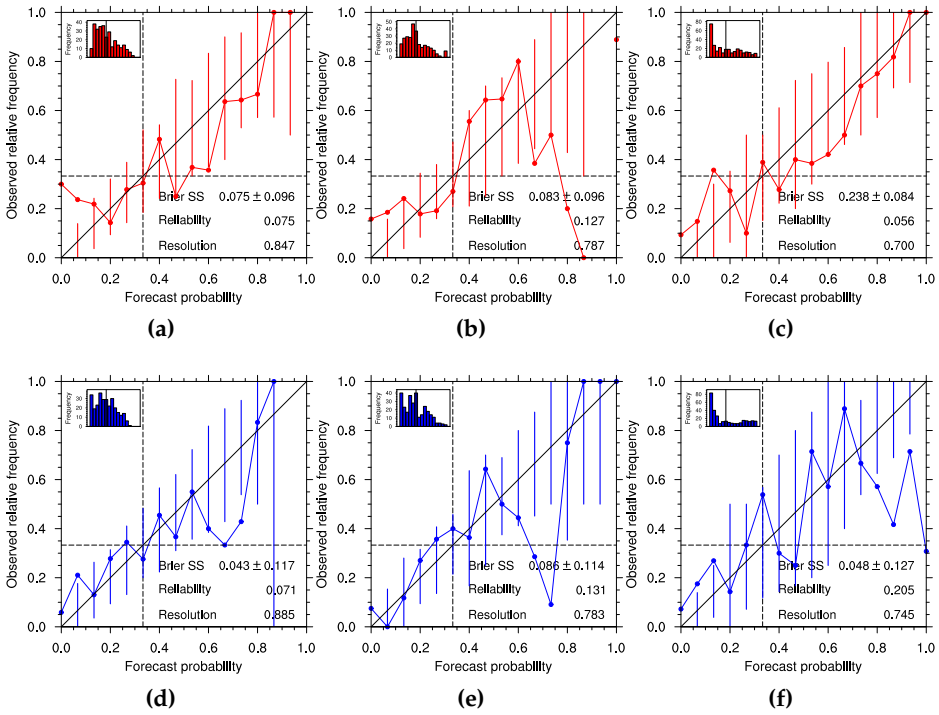


Figure 6.25: Reliability of upper (a-c) and lower (d-f) tercile DJF temperature forecasts over Malawi. Reliability is shown for System 4 forecasts issued August (a & d), October (b & e) and December (c & f).

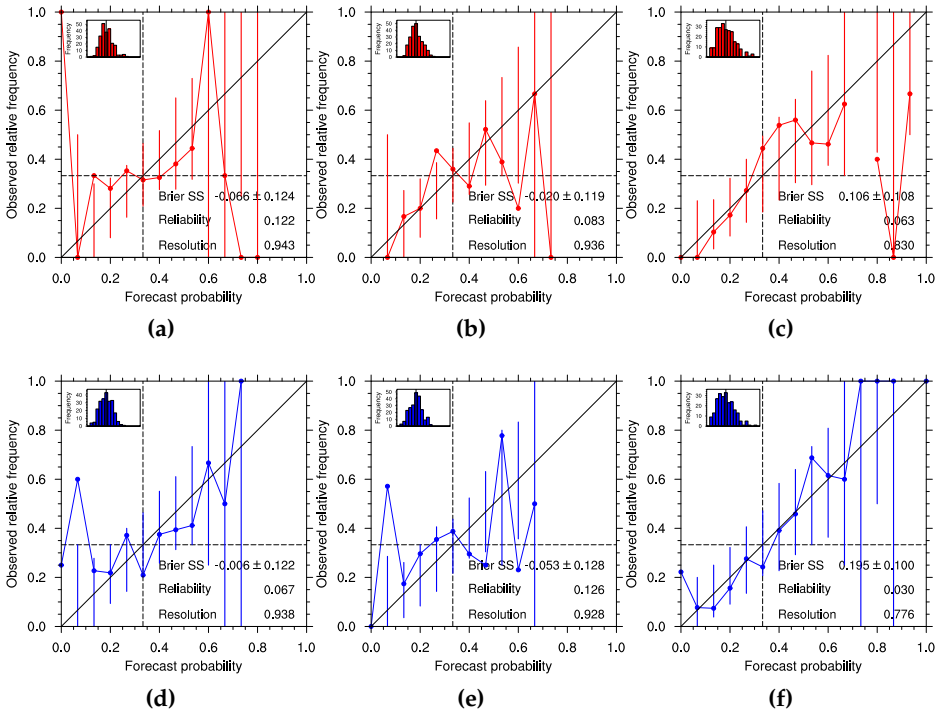


Figure 6.26: Reliability of Malawi precipitation vs GPCP, details as in figure 6.25.

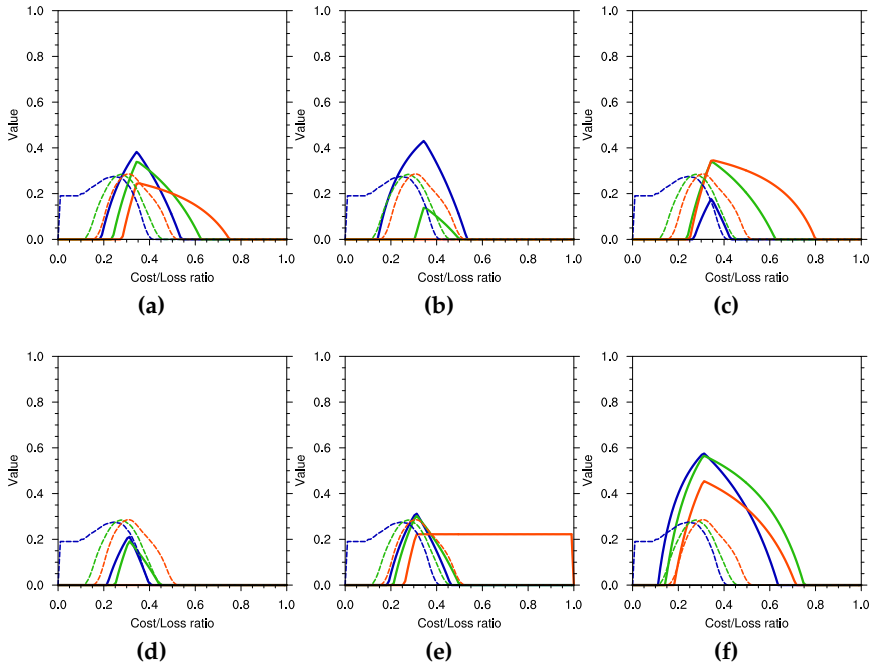


Figure 6.27: Value of upper (a-c) and lower (d-f) tercile JAS temperature forecasts over Malawi, for System 4 forecasts issued August (a & d), October (b & e) and December (c & f). Curves for 30%, 50% and 70% decision thresholds are represented by blue, green and orange lines with dashed lines indicating 95% significance level for each threshold.

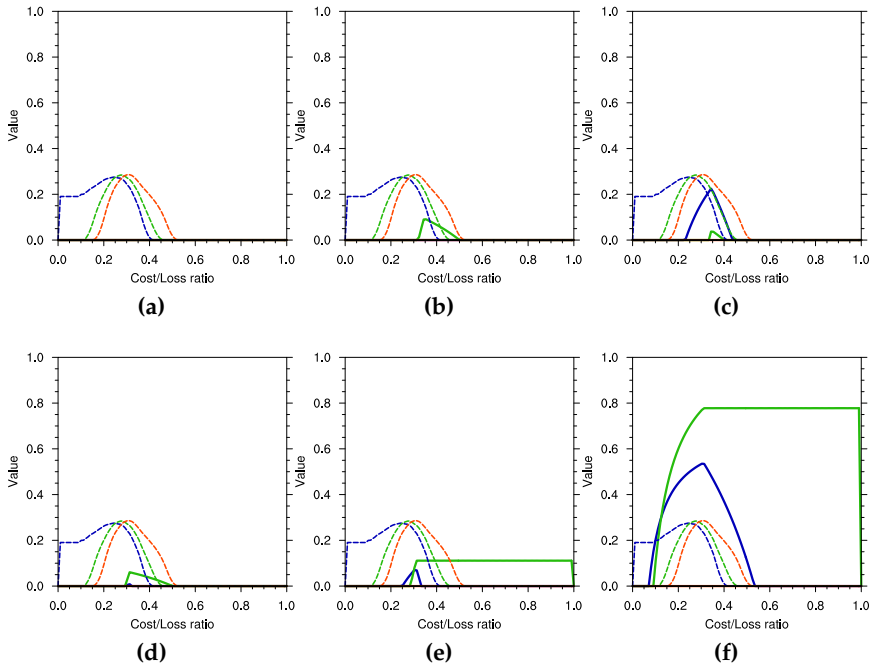


Figure 6.28: Value of Malawi precipitation vs GPCP, details as in figure 6.27.

6.2.3 Interpreting the different metrics

There are multiple metrics to measure the behaviour of a forecast system over a hindcast period. Several have been considered whilst looking at seasonal prediction systems: in this short section they are compared, and consideration is given to how information contained in economic value curves might translate to communication to decision makers and operational use.

Figure 6.29 shows the bias, ensemble mean correlation and ROC AUC for JAS precipitation forecasts over West Africa (upper tercile in the case of the ROC AUC), for forecasts issued in March and July. These forecasts were selected as skill (particularly economic value) is higher closer to the target. There is a similiarity between the plots: particularly between ensemble mean correlation and ROC AUC; regions with high and low scores are the same according to both metrics. Similarly, for the earlier lead time where the scores are lower, the bias is higher, and when the scores are higher closer to the target, the bias is much reduced.

This is the general behaviour one might expect from a forecast system: if it is simulating reality well metrics would tend to give it higher scores, whilst if it is a poor simulation scores should be accordingly low. Different scores give different information however: biases can give clues as to the mechanistic reasons for poor performance (looking at biases in other fields, e.g. surface pressure and winds would be necessary to further investigate source of prediction error). Looking at biases is also useful from the forecaster/impact modeller perspective, as they indicate what post-processing may be necessary for further use of model output. Ensemble mean correlation and ROC AUC both give an idea of the skill in prediction: correlations an overall view, whilst ROC AUC is specific to a certain event.

Figure 6.30 shows the reliability diagrams, forecast probability bar charts and economic value curves for the same forecasts in figure 6.29. For the forecast probability bar charts the decision trigger threshold chosen to calculate hits, misses, false alarms and correct rejections is the one from 30, 50 and 70% which gives the largest integral under the value curve. In the figure, observed events are indicated by the black bars, whilst white bars signify years when the event did not occur.

The reliability diagram suggests that the March forecast is more reliable than the July forecast, since the curve for March lies well within the consistency bars for most points, whilst in July it is only within the bars for a few points. However the forecast barcharts and value tell a different story: forecasts have much more value in July than they do in March. A breakdown of the source of this value follows.

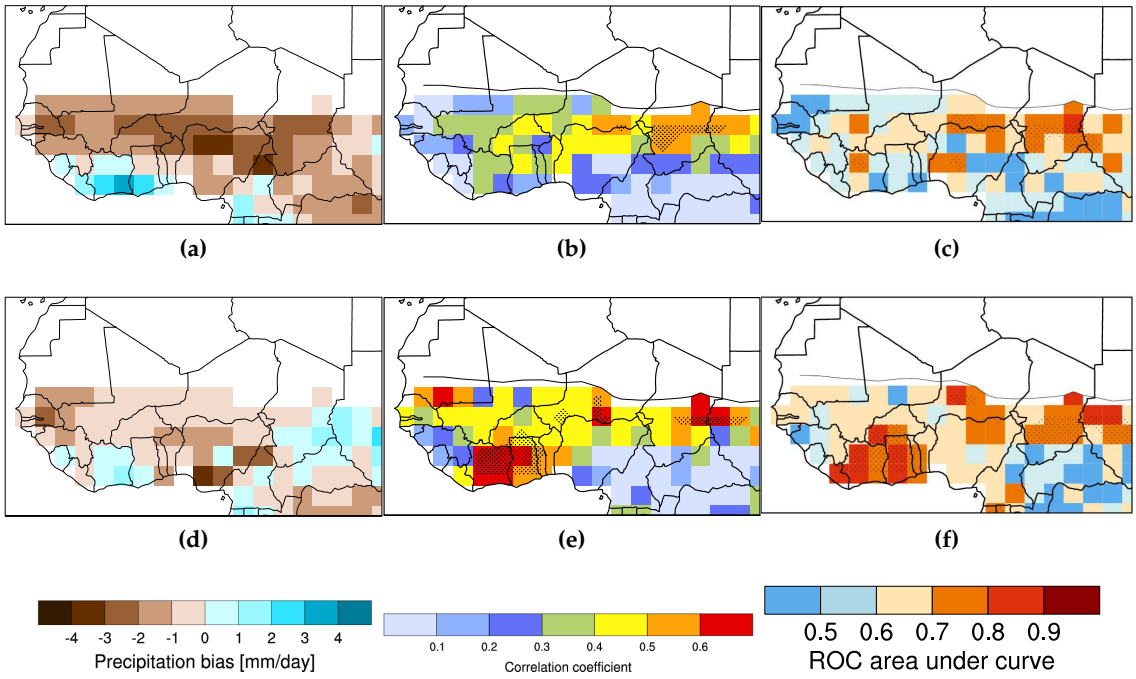


Figure 6.29: A collation of (a & d) bias, (b & e) ensemble mean correlation and (c & f) upper tercile ROC AUC for West Africa JAS precipitation. From 'poor' forecasts issued in March (top row) and 'good' forecasts issued July (bottom row). N.B. These plots are repeats of those previously appearing in the text; further details can be found in the corresponding captions.

In March, the system has value by virtue of correctly identifying one event. If a decision maker acted on this information, a loss would be avoided. However, every other event would be missed. This information then is only valuable if it is normally cheapest to never take action and just experience losses. If it is cheapest to act all the time then following the forecast issued in March gives no value, as it would only advise action once, and cause the loss associated with nine events. In this case it would be likely to be cheaper in the long run to ignore the forecast and always take preventative action: i.e. is it is better to always be safe than sorry when being safe is cheap and being sorry is expensive.

For July forecasts, the number of hits is much higher and the number of false alarms and misses is low. There is value in using the forecast across a wide range of cost/loss ratio, i.e. if normal behaviour is either to always or to never act. If the cost/loss ratio is high and one would normally never act, the forecast saves money by spending a large amount of money on prevention and successfully reducing losses, with only one instance of unnecessary expenditure (one false alarm). If instead the cost/loss ratio is low and one would normally always act, the value is obtained by successfully knowing when not to act and saving some money on prevention, with only two misses where the money saved

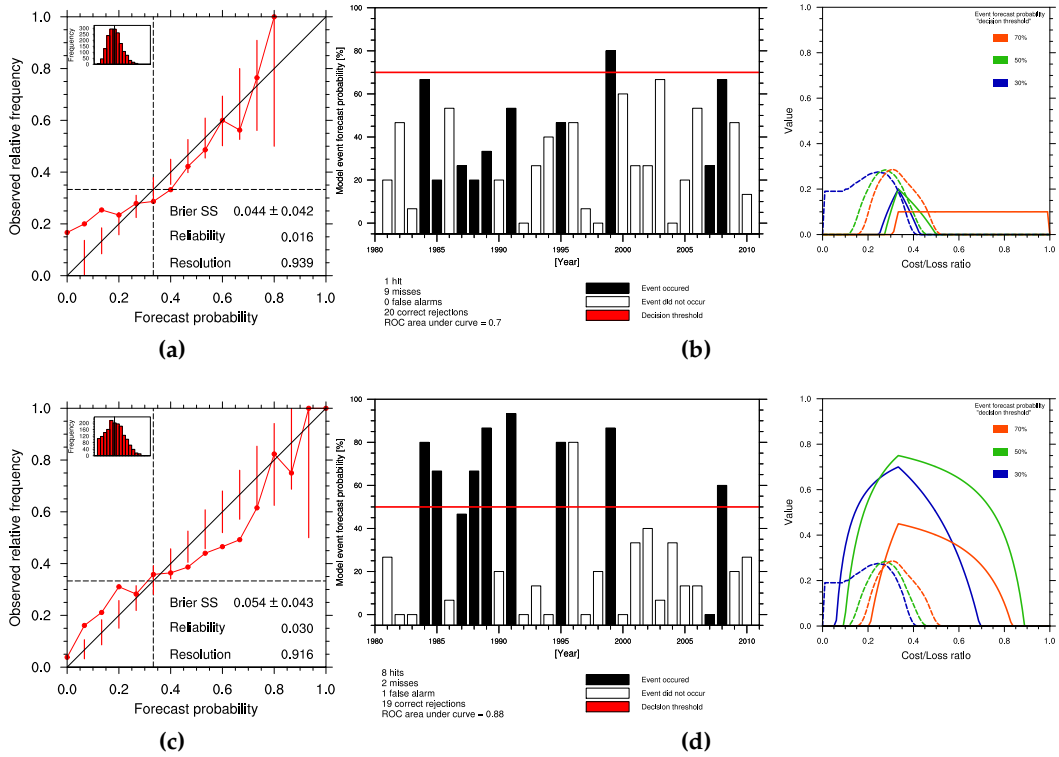


Figure 6.30: Reliability (a & c) and forecast probabilities (b& d) for upper tercile JAS precipitation over the Gulf of Guinea. The corresponding potential economic value is shown alongisde the forecast probabilities. From 'poor' forecasts issued in March (top row) and 'good' forecasts issued July (bottom row). For the forecast probabilities, GPCP upper tercile precipitation events are signified by black bars; non-events are white. Bar height indicates System 4 forecast probability for an upper tercile event. Hits, misses, false alarms and correct rejections are defined using a threshold chosen separately for each start date, based on a visual estimation of the threshold which maximises the area under the value curve.

by not acting is reduced by the expense of losses. This value across the cost/loss domain is reflected in the value plot where the value curve is positive for any situation when the cost of acting is between 10% and 90% of the expense of the loss.

So how is it that the March forecasts are more reliable but have less value than the July forecasts? Reliability, as measured by the reliability curve, is defined by the observed frequencies conditioned on forecast probabilities, whilst one economic value curve is defined for a single forecast probability. For the reliability curve each point is defined only by a subset of the forecast-observation pairs falling inside a single bin, whilst for economic value the whole curve depends on the entire set of forecast-observation pairs. There may then be a 'sweet spot' in the forecast probabilities which gives good economic value, where outside this point the forecast system is poor. Put another way, a

forecast system showing a poor reliability curve may still provide economic value, since the probability information is reduced to a binary indicator (act/don't act) across the forecast period. One may then not trust the explicit forecast probabilities from the model (informed by the reliability diagram), yet be able to extract decision-useful information from it by calibrating a decision to a specific threshold (informed by potential economic value curves).

This may be illustrated by an extreme example. Consider a forecast system issuing equally distributed probabilities of an event, between 0 and 100%. Now say that across the forecast-observation pairs, every time the probability is below 50% the event does not occur and every time it is over 50% it does. From the perspective of economic value, choosing a threshold of 50% makes the forecast system perfect: hitting every single event and correctly rejecting every non-event. However if one constructed a reliability diagram for this forecast system based on discretising the forecast probabilities into bins, no single point except for the corners (0,0) and (1,1) would lie on the diagonal. All bins below 50% (e.g. 0-10%, 10-20% etc.) would have an average of 0 for their corresponding observations, whilst those above 50% (50-60% and so on) would have an average of 1. This would create a reliability diagram looking like a step function. Yet the performance of the forecasting system when calibrated based on economic value would be no different from a perfect system. This shows therefore that no metric is perfect and each is useful for a purpose. If one needs to know if the exact probabilities offered by the model can be trusted then the reliability curve is necessary. However poor reliability does not necessarily mean a useless forecast: by looking at potential economic value, unreliable forecast systems may be redeemed.

Conversely of course it is possible to have a perfectly reliable yet valueless forecast system: consider climatology. Since every forecast probability is by definition the event frequency, the observed frequency conditioned on the forecast probabilities is just the event frequency, and the reliability diagram is a single point, lying on the diagonal. That is, the system is perfectly reliable. It is however, useless, as there is (by definition) no reduction of expense compared to climatology.

6.3 Discussion: when can useful forecasts be made?

The analysis in this chapter shows how the System 4 hindcasts would have performed if were they been available over the period 1981-2010. For the regions studied it can answer the question: 'If this forecast were available, would it have been useful?'.

Only a subset of all System 4 start dates have been discussed, for reasons of brevity. System 4 forecasts run for seven months ahead of initialisation, so for a three month season, five potential forecasts are available, the first forecast initialised four months ahead of the target and the last initialised at the start of the target season. A summary of the economic value of forecasts at all lead times (including those not discussed in the text) is given in table 5.5.

It can be seen from this table, and from the description of results in the previous section, that skill and its evolution as a target is approached is variable. Generally the skill of temperature forecasts is higher than precipitation, as is normally the case with climate predictions (e.g. chapter 4). For some regions the economic value of forecasts is high, for example in the Sahel and the Gulf of Guinea for temperature and precipitation. Other regions do not have high skill at any lead time (e.g. temperature predictions for Bangladesh).

Plots of ROC AUC can highlight the subregions not studied in this analysis where skilful predictions may be made. Example regions are a subset of the Sahel for JAS precipitation (figure 6.6), DJF precipitation over Tanzania (figure 6.20) or MAM precipitation over Uganda (figure C.6). Further analysis based on these plots could lead to useful climate predictions for these areas.

ROC areas and ensemble mean correlations generally seem to be well correlated. In areas of low correlation ROC AUC is low, whilst points of significant ROC AUC invariably have high correlations. Differences in skill between start dates are mirrored in the ensemble mean correlation and the ROC AUC. This is not surprising, since in a region where a forecast is simulating reality well (i.e. where the paths of forecast ensemble members in 'climate-space' track close to reality) one might reasonably expect a high correlation of ensemble mean with observations as well as expecting a good verification of upper and lower tercile events.

Does this mean that there is degeneracy in using both verifications? Perhaps some; however there is certainly more information in ROC AUC than in ensemble mean correlation. For instance there is often a difference in skill between upper and lower tercile forecasts (for example compare upper and lower tercile October forecasts of DJF precipitation over Botswana in figure 6.20b). It is arguable that there is no extra

			Lead time (months)				
Variable	**Region-Season**	**Event**	4	3	2	1	0
Temperature	Sahel JAS	UT					
		LT					
	Gulf of Guinea JAS	UT					
		LT					
	Botswana DJF	UT					
		LT					
	Malawi DJF	UT					
		LT					
	Kenya MAM	UT					
		LT					
	West India JJA	UT					
		LT					
	Bangladesh JJA	UT					
		LT					
Precipitation	Sahel JAS	UT					
		LT					
	Gulf of Guinea JAS	UT					
		LT					
	Botswana DJF	UT					
		LT					
	Malawi DJF	UT					
		LT					
	Kenya MAM	UT					
		LT					
	West India JJA	UT					
		LT					
	Bangladesh JJA	UT					
		LT					

Table 6.2: Summary of potential economic value of upper and lower tercile average temperature and precipitation forecasts for all regions defined in figure 5.1, upper and lower tercile forecasts. Results are shown for all lead times. Colours indicate the magnitude of the value in each case: red for when there is no value for any threshold above significance, orange where the value is just above significance and yellow where there is value clearly above significance yet not more than 0.5 for any threshold. Finally, green indicates where at least one curve has a value above 0.5.

information in ensemble mean correlation maps that is not in maps of ROC AUC (compare the October Botswana precipitation ROC AUC with the corresponding correlation map in figure 6.18b).

Despite this degeneracy both measures of skill are still useful. Recalling the 'progressive disclosure of information' discussed in section 2.2.3; where uncertainty information is

> *progressively disclosed, from non-technical information through more specialised information according to the needs of the user* (Pereira and Quintana, 2002),

it is easy to envisage a situation where an ensemble mean correlation map may be the best way to communicate model skill. To start with ROC AUC maps for both upper and lower tercile events (as well as explaining the meaning of the less-familiar metric), may be to overload a non-specialist. Whilst there is some information overlap between the two scores, the ROC AUC is perhaps better left to a later stage of dialogue with users, in order to ensure effective communication.

It is important to explore all aspects of a forecast before using it or giving to a decision maker, and to tailor the message to the purpose of the dialogue. Thus, an initial interaction can begin with simple metrics summarising skill, moving through more complex scores like ROC AUC and value to an explicit visualisation of how the forecast system for an individual region would have performed over the whole hindcast period, such as shown in figure **??**.

One might expect that forecasts initialised closest in time to a target will make better predictions than those initialised further away, but this is not always the case. For some forecasts this behaviour is observed, where four month lead forecasts are poorest as measured by the value (see table 6.2), with value increasing as the target is approached (e.g. Gulf of Guinea UT temperature, Sahel LT precipitation and West India UT precipitation). However there are other regions where the value is either constant with time (e.g. Malawi UT temperature), or even where the forecast issued closest to the target has the lowest skill (e.g. Sahel UT temperature, Malawi UT precipitation).

When should a forecast be issued to a decision maker? The answer to this depends on a user's cost-loss ratio and the relevant value plot should be studied with a specific decision in mind. In some places there is no economic value at any lead time, and so forecasts are then not useful for decision makers (e.g. West India JJA precipitation figure C.20). This information is still useful to modellers, since it can guide detailed study into the regional climate dynamics to deduce the source of forecast error and allow model improvement.

Where there is potential economic value, when exactly a forecast should be issued

depends on the nature of the decision to be made. A decision may one-shot decision; for example if a humanitarian agency needs to choose if to launch an emergency appeal in advance of flooding (as in (Tall et al., 2012). In this case a decision cannot be reversed and so a user may want to wait for the start date when value normally saturates at a high level (e.g. waiting for the one month lead for lower tercile precipitation over the Gulf of Guinea, see figure 6.14 and also table 5.5).

On the other hand, preventative action can evolve over time, such as in the case of gradual resource and emergency supply distribution. In this situation a forecast with evolving skill may be useful, for long-lead action where the loss associated with an incorrect forecast is minimal. A forecast of this nature might be for lower tercile DJF precipitation over Malawi (figure 6.28 and table 5.5). Here a forecast at three months ahead could be provide some guidance, whilst action can still be modified when forecasts improve at the start of the rainy season.

However as mentioned above, in some places forecasts get worse as a target is approached and the reason for this is not clear. This could potentially be perhaps due to initialisation. A speculation might be that there may be issues with reanalysis for certain periods of the hindcast period. That is, if there is a bias in July SST near a target region, then forecasts initialised with this data will include this error in their JAS prediction. If the reanalysis for previous months does not have a bias, then a March forecasts could potentially have true JAS skill, whilst the July forecast does not. This seems unlikely however, as it would require error in the initialisation data for the same month for several years of the hindcast period. Another potential reason for the observed effect is dynamical; perhaps initialising a forecast once the rains have started creates error in the prediction. However, no satisfying reason for this worsening of forecasts has been determined.

For regions where forecasts worsen as a target is approached and the source of the error cannot be determined, should a decision maker use the prediction? Certainly it may introduce some uncertainty into a forecast, reducing trust (in the way that a skilful black box prediction is less trustworthy than a model where the source of predictability can be explicitly demonstrated). But should a March forecast for JAS precipitation be issued if we know that the model cannot predict JAS precipitation when initialised in July? There is no obvious answer to this question and if it is decided to only issue this early forecast, any reservations climate modellers have about it should be communicated to users.

It has been shown that climate predictions are imperfect: at decadal scales predictions do not have sufficient skill for impact prediction (chapter 4), whilst the skill of seasonal climate predictions is variable in time and space (chapters 5 and 6). The relationship between climate variables and disease is non-linear, and a direct map from seasonal

average climate to disease outcomes does not exist. However, it is possible to link a disease model with seasonal climate model output; how successfully disease predictions made by this method are is the subject of part two of this thesis, along with an exploration of the uncertainty associated with doing so.

Part II

Using climate forecasts to predict disease risk

CHAPTER 7

The skill of dynamical seasonal climate-driven malaria forecasts

This chapter contains validation of malaria forecasts driven by ECMWF System 4. There is a link between climate and malaria, and by using seasonal climate predictions it is theoretically possible to make malaria risk forecasts. However, using output from one model as an input to another increases complexity and uncertainty and models must be carefully validated before using them to provide early warnings. Here forecasts of malaria incidence are validated at a tier 3 level over Botswana and at tier-2 level over the Sahel, the Gulf of Guinea and Malawi.

It has been shown that it is possible to make good predictions of seasonal average temperature and precipitation during the rainy season (for some regions, under certain conditions: see chapter 6). Temperature and precipitation are important variables for disease prediction and so the skill of a climate model for rainy season average prediction enables a pre-selection of which region to study: if the average conditions during the rainy season cannot be predicted in a particular region, it is likely that a climate-driven disease model will fail to provide good disease forecasts. Conversely, a place for which good seasonal climate predictions can be made is a good candidate for making disease forecasts[1].

By using the Liverpool Malaria Model (LMM; Hoshen and Morse, 2004), a process-based model for seasonal malaria driven by daily temperature and rainfall, the potential for

[1] N.B. Skilful prediction of seasonal totals is not a necessary nor sufficient criteria for making good disease predictions, since it is possible that properties of the climate beyond average conditions (e.g. number of rainy days, timing of the rainfall peak) may provide a strong source of predictability. However the ability to successfully predict average conditions suggests a somewhat realistic climate model and is sufficient information to prioritise a region for further study over other regions for which predictions of average conditions are poor.

disease prediction can be explored. Direct, tier-3, validation of LMM output is possible for Botswana, where observations are available in the form of a standardized index of cases of laboratory-confirmed malaria incidence for the period 1982-2003 (Thomson et al., 2005). This index has been previously used to validate the LMM at tier-3 level, using the DEMETER seasonal hindcasts (Thomson et al., 2006) and here malaria forecasts for Botswana are carried out using System 4 to drive the LMM.

Elsewhere tier-3 validation is limited by observational malaria data constraints (as discussed in section 3.2). Despite this, tier-2 validation can be performed, using the LMM driven by climate observations as a target. This is carried out here for a selection of regions where System 4 climate predictions show value; over the Sahel, the Gulf of Guinea and Malawi. For these regions the System 4-driven LMM is validated against the LMM driven by the ERA-Interim reanalysis.

The following methodology section contains a short description of the LMM and details relating to using the LMM with both the System 4 hindcasts and the ERA-Interim reanalysis. Results after this are divided into three sections; firstly results for direct validation over Botswana are described, following this are results for tier-2 validation over other regions. The final section of the results suggests a method of using tier-2 and tier-3 validation to quantify the uncertainty in climate-driven disease risk forecasts, and the chapter ends with a general discussion.

7.1 Methodology

7.1.1 The Liverpool Malaria Model

Full details of the LMM are given elsewhere (Hoshen and Morse, 2004; Morse et al., 2005), a short description of the climate components follows here.

The LMM uses a dynamic approach to simulate malaria incidence in the human population, and consists of two climate driven components. The first is the mosquito population component, which is modelled using larval and adult stages. In the model the number of eggs deposited into breeding sites depends on the previous ten days' rainfall, as does larval mortality rate. The adult mosquito mortality rate depends on temperature.

The second component is the process of parasite transmission between human and mosquito hosts. There is a temperature dependency in the gonotrophic and sporogonic

cycles and in the mosquito biting rate[2]. Both the gonotrophic and sporogonic cycles progress at a rate dependent on the number of 'degree days' above a specific temperature threshold: the gonotrophic cycle takes 37 degree days above a threshold of 9°C and the sporogonic cycle takes 111 degree days with a threshold of 18°C. This latter threshold is one of the most critical areas of sensitivity in the model; below it no parasite development can occur.

7.1.2 Tier-3 validation over Botswana

The malaria index for Botswana is a time series of cases of laboratory-confirmed malaria incidence for January to May over Botswana for 1982-2003, converted to standardized anomalies (Thomson et al., 2005). This is used as a target for tier-3 validation and is hereafter referred to as the Botswana Malaria Index (BMI). For validation against this index the LMM was driven by System 4 over Botswana (as described in the following section), using forecasts initialised in December and earlier. The resultant output was then averaged temporally across Jan-May and spatially over Botswana (defined as 17.5-27.5°S and 19.5-29°E). Finally, standardized anomalies have been calculated (for each start date separately) for the same time period, with the same done for LMM driven by ERA-Interim. Results are presented in the form of boxplots showing the 5-25-50-75-95th percentiles of the forecast ensemble, along with the ROC AUC for the area average. Results for upper and lower tercile forecasts are summarised, and the potential economic value of the forecasts is also presented.

7.1.3 Driving the LMM with System 4 and ERA-Interim

Previous work validating the LMM at a tier-2 level used the LMM driven by the ERA-40 reanalysis as a target (Jones and Morse, 2010). Here the updated ERA-Interim reanalysis product was used. ERA-Interim is a daily gridded dataset spanning the whole hindcast period of System 4 and of the reanalysis products it has the closest spatial resolution to System 4 (further description of ERA-Interim is given in 3.1).

The forecasts studied are those which make a prediction for a three month rainy season. Since System 4 hindcasts make forecasts for seven months from initialisation, a full three month climate can be predicted using System 4 up to four months in advance, giving five possible start dates (including one initialised at the start of the season). In each case

[2]Here gonotrophic cycle describes the process of blood-feeding, egg maturation and ovipositioning, repeated several times throughout a mosquito's life cycle, whilst the sporogonic cycle refers to the development of the *Plasmodium* parasite within the mosquito.

an ensemble of 15 members is available for each start date, and these are interpolated to the lower-resolution ERA-Interim spatial grid.

Before driving the LMM with the System 4 hindcasts, it was spun-up with a 366 day smoothed System 4 climatology, each time starting at the same start date as the corresponding hindcast. Advice from other users of the malaria model (Volker Emert, personal communication), was that the model requires multiple years of spin-up climatology. This was tested, and whilst there was a difference between using no spin-up and one year of spin-up, to use any more years than this did not make any apparent change in the results. Consequently only one year of the seasonal cycle was used as a spin-up. When run with the ERA-Interim reanalysis a 366 day ERA-Interim climatology was used.

The employed target regions for tier-2 validation are the Sahel, the Gulf of Guinea and Malawi, as previously defined in figure 5.1. The choice is based on the skill highlighted in a tier-1 validation context as demonstrated in chapters 5 and 6; in these regions System 4 forecasts have value for temperature and precipitation (see table 6.2). Three month seasons with high rainfall are taken to be July-August-September (JAS) for the West African regions and December-January-February (DJF) for Malawi. The System 4 start dates used for each region are those which fully contain the rainfall season. The malaria season generally falls a few months subsequently to the rainfall season, which has a slight variation across regions, and by looking at the malaria climates this corresponds to malaria seasons occurring in September-October-November (SON) for the Sahel and the Gulf of Guinea, and March-April-May (MAM) for Malawi.

For longer lead times, the System 4 reforecast ends before the end of the target malaria season (due to the lag between climate and malaria). To deal with this issue, the remaining time was filled with System 4 climatology. Whilst it is unlikely, it may be the case that there is some skill in the temperature anomaly at the end of month seven. To exploit this, the average temperature of the final week of the forecast is persisted into the first month after the end of the reforecast. This anomaly is relaxed to the 366 day climatology, using an exponential decay function with a decay constant of 1/10 days. This corresponds to a reduction to roughly 1/3 of its initial value after 10 days and a near complete return to climatology by the end of the month.

For precipitation no anomaly is persisted and the field is replaced by the daily climatology immediately after the forecast ends. This is due to both the questionable realism of allowing a precipitation anomaly to persist beyond the main months of the rainy season, and also the unlikelihood that precipitation forecasts are still skilful at seven months lead time.

A simple correction of temperature bias was also carried out; a daily bias was calculated by simply subtracting the (smoothed) 366 day climatology ERA-I 366 from System 4. This 366 day bias was then subtracted from individual year hindcasts before running through the LMM. No bias correction of precipitation was attempted.

7.2 Results

Results are summarised in table 7.1. Following this results are described in detail for tier-3 validation over Botswana and then tier-2 validation over the Sahel, the Gulf of Guinea and Malawi. In each case shown three start dates are presented: at the start of the rainy season, and for the two preceding months. Where there is no malaria data (i.e. at the tier-2 level), more extensive exploration of the behaviour of the forecasts is presented: looking at driving climates and spatial maps as well as simulated incidence, coefficient of variation and ROC AUC maps. The final section of the results describes a method of interpreting validation as a quantification of uncertainty.

Validation level	Region: Target	Main Results
Tier-3	Botswana: Jan-May	• Significant ROC AUC and potential economic value for forecasts initialised November • Some skill for upper tercile forecasts initialised December after temperature bias correction • No skill for lower tercile forecasts initialised December, nor for October start dates
Tier-2	Sahel: Sep-Nov	• Some significant ROC AUC at the epidemic fringe in the Western Sahel for forecasts initialised July, below significance outside this region • No skill for forecasts initialised May or June • Temperature bias correction does not increase ROC AUC
	Gulf of Guinea: Sep-Nov	• ROC AUC below significance over entire region for all start dates • Temperature bias correction gives no improvement
	Malawi: Mar-May	• Skill over the north west for start dates in October, November and December; no significant scores outside this region • Temperature bias correction does not improve forecasts

Table 7.1: Summary of malaria incidence forecasts using the LMM driven by System 4.

7.2.1 Tier-3 validation over Botswana

ROC AUC over Botswana for January to May average malaria incidence are summarised in table 7.2 for System 4 start dates in October, November and December.

		Raw			T bias correct		
Reference	Event	Oct	Nov	Dec	Oct	Nov	Dec
BMI	UT	0.36	**0.86**	0.55	0.47	**0.85**	**0.78**
	LT	0.30	**0.80**	0.65	0.51	**0.80**	0.67
ERA-I	UT	0.50	0.69	0.65	0.63	**0.81**	**0.90**
	LT	0.52	0.72	**0.87**	0.61	**0.78**	**0.85**

Table 7.2: Summary of ROC AUC for tier-3 validation over Botswana. Bold numbers are where ROC AUC is different from climatology at 95% significance (based on a comparison with the Mann-Whitney U test; see chapter 3).

Measured against the BMI the score for forecasts issued in October is not significantly different from climatology at the 95% level, whilst November forecasts are significant before and after bias correction, and System 4 December forecasts are only significant after temperature bias correction for upper tercile events only. When using ERA-Interim as a reference, only lower tercile forecasts issued in December are significant before bias correction, whilst bias correction improves the score so that ROC AUC is significant for both upper and lower tercile categories for November and December start dates.

Boxplots of the System 4 ensemble forecast for January-May incidence are shown in figure 7.1, for non-bias corrected November start dates (where the ROC AUC is highest according to table 7.2). The BMI and the ERA-Interim driven forecasts do not exactly match (Pearson's rank correlation coefficient between the two time series is 0.65). For some years the definition of tercile events by the two datasets is the same (e.g. 1982 and 1988), whilst for other years it disagrees (e.g. 1990 and 1993).

Potential economic value plots for upper and lower tercile events for October, November and December start dates are shown in figure 7.2 for raw forecasts and 7.3 for forecasts with a temperature bias correction, using the BMI as a reference. Results here are consistent with the ROC AUC scores; no value is observed for October start dates, whilst the highest value is seen for November start dates. Bias correction increases the value of December forecasts for the upper tercile category whilst lower tercile forecasts in December do not have ROC AUC above significance when measured against the BMI, though they do when measured against ERA-Interim driven hindcasts.

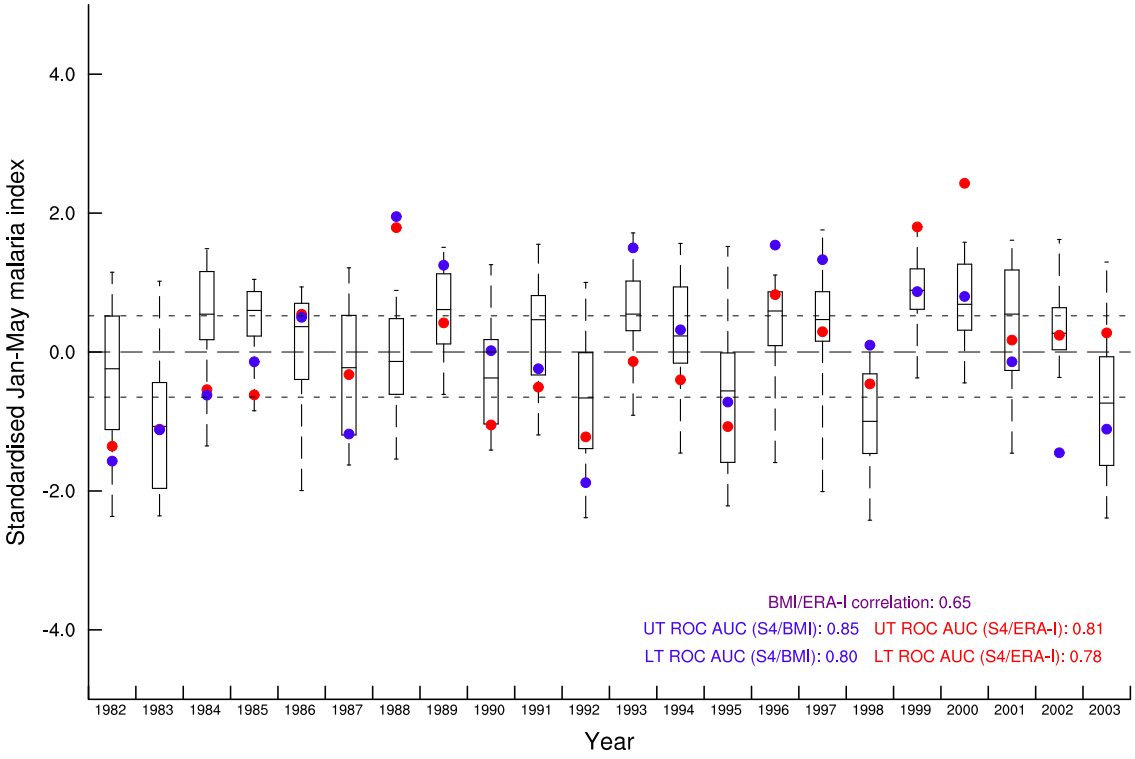


Figure 7.1: Standardised System 4 malaria forecasts over Botswana for 1982-2003, forecasts issued in November. Boxes show 5th, 25th, 50th, 75th and 95th percentiles within each forecast ensemble. Blue dots indicate the Botswana Malaria Index (BMI, Thomson et al., 2005) value for that year and red dots indicate the corresponding value of the LMM driven by ERA-Interim. Dashed lines indicate upper and lower tercile boundaries of the BMI data.

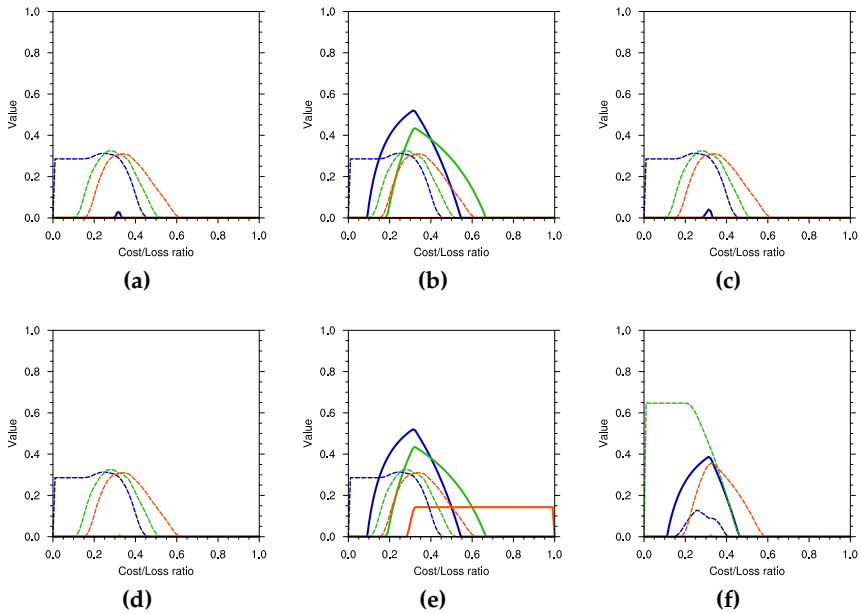


Figure 7.2: Value of upper (a-c) and lower (d-f) tercile malaria incidence forecasts for Jan-May over Botswana measured against the BMI. Value vs. cost/lost ratio is shown for System 4 forecasts issued in October (a & d), November (b & e) and December (c & f). Curves for 30%, 50% and 70% decision thresholds are shown by blue, green and orange lines, with dashed lines indicating 95% significance level for each threshold. No bias correction has been carried out.

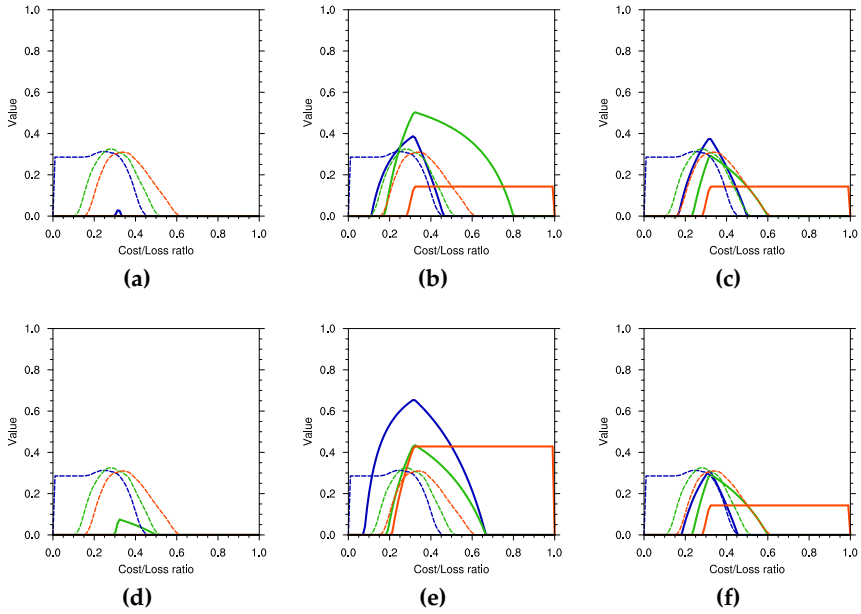


Figure 7.3: As figure 7.2, for temperature bias corrected malaria forecasts

.

7.2.2 Tier-2 validation over African regions

Tier-2 validation is now described for African regions beyond Botswana. Results follow for Botswana, the Sahel, the Gulf of Guinea and Malawi, beginning in each case with an analysis of the System 4 simulation of seasonal cycles of climate drivers and malaria incidence, followed with results relating to the mean malaria climate; mean, standard deviation and the coefficient of variation (COV). COV here is defined simply as the division of the mean by the standard deviation, giving a measure of the magnitude of a variable with respect to its variability. Finally maps of ROC AUC for malaria incidence are presented. Some figures are discussed in the text though are left to appendix D, to reduce the number of figures in the main body of description.

Botswana

Whilst tier-3 validation for Botswana is possible and has been carried out in the preceding section, it is illuminating to consider tier-2 validation for the same region. This allows an examination of the link between tier-2 and tier-3 validation.

Climatologies for precipitation and temperature are shown in figure 7.4. The precipitation cycle is unimodal, with a peak in December to February; outside this time the rainfall is low, dropping to almost zero in June to August. System 4 over predicts precipitation, by around one mm/day in the rainy season, though the shape is correct. Temperature follows the same cycle as precipitation, with the warmest time of year coinciding with the rains. System 4 is slightly too cold, which may be associated with the excess precipitation.

Maps for precipitation are shown in appendix D, in figure 7.4a for ERA-Interim and 7.4b for System 4. The pattern of precipitation is generally correct, with the rains beginning and ending at the right time of year. During the rainy season there is more rainfall in the north and south east of the region, which is well simulated in System 4 (though with a wet bias). For the temperature, shown in figure D.3, there is a stronger bias during the rainy season, which occurs mainly in the south west of the region. This is in the place where rainfall is lowest, suggesting that the bias is not due to excess rain and instead related to incorrect radiation balance through absence of cloud.

For malaria, climatology is shown in figure 7.4e for ERA-Interim and 7.4f for System 4. The shape is unimodal, with incidence above 40% between February and May only, peaking at around March. The climatology is similar for System 4 and ERA-Interim, except for an over simulation of malaria in February. This is likely due to the wet bias in the rainy season. Maps for incidence climatology are shown in figures D.4 and figure D.5,

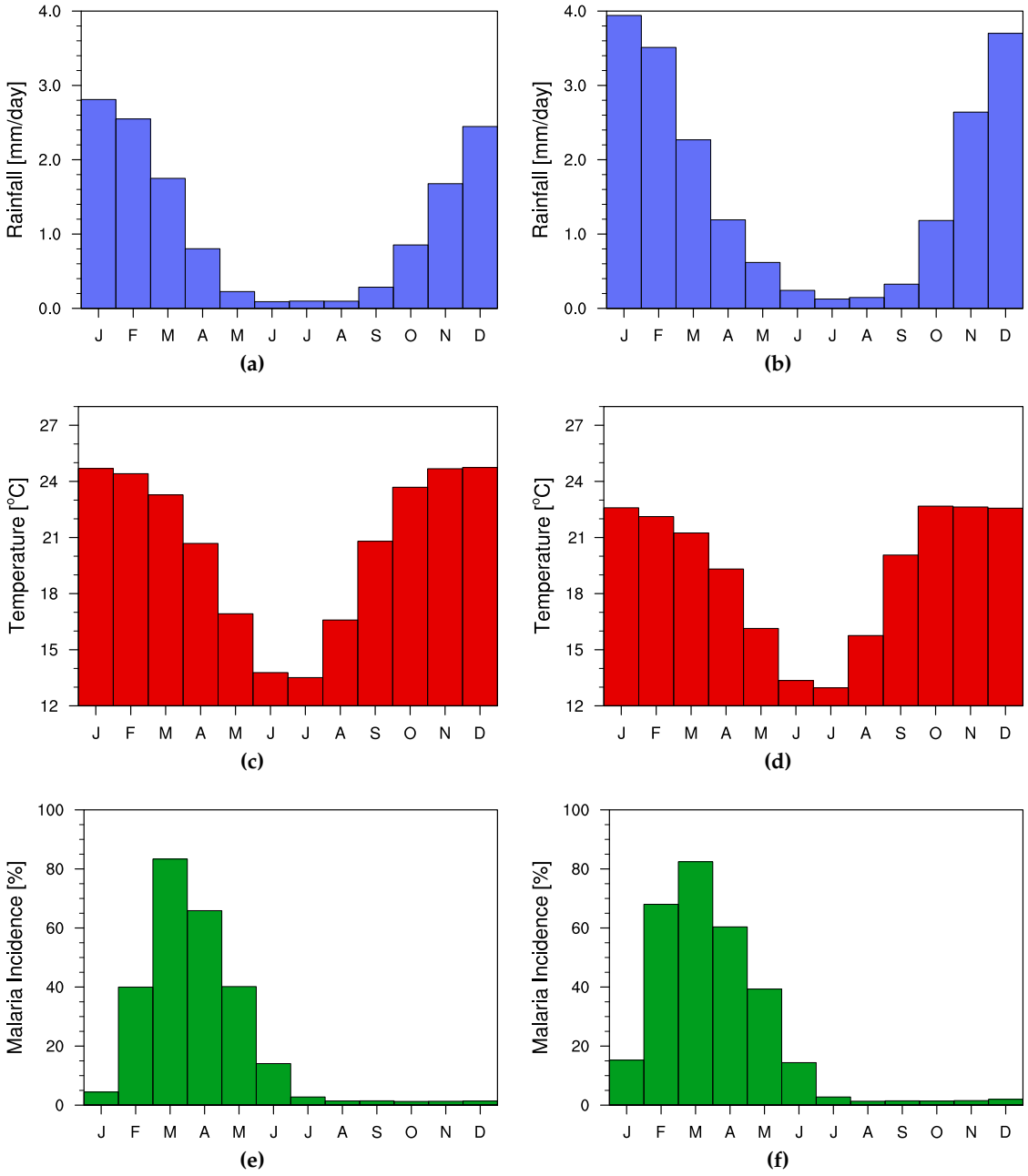


Figure 7.4: Precipitation (a, b) temperature (c, d) climatologies and LMM-simulated incidence for ERA-Interim (left column) and System 4 (right column) for Botswana

.

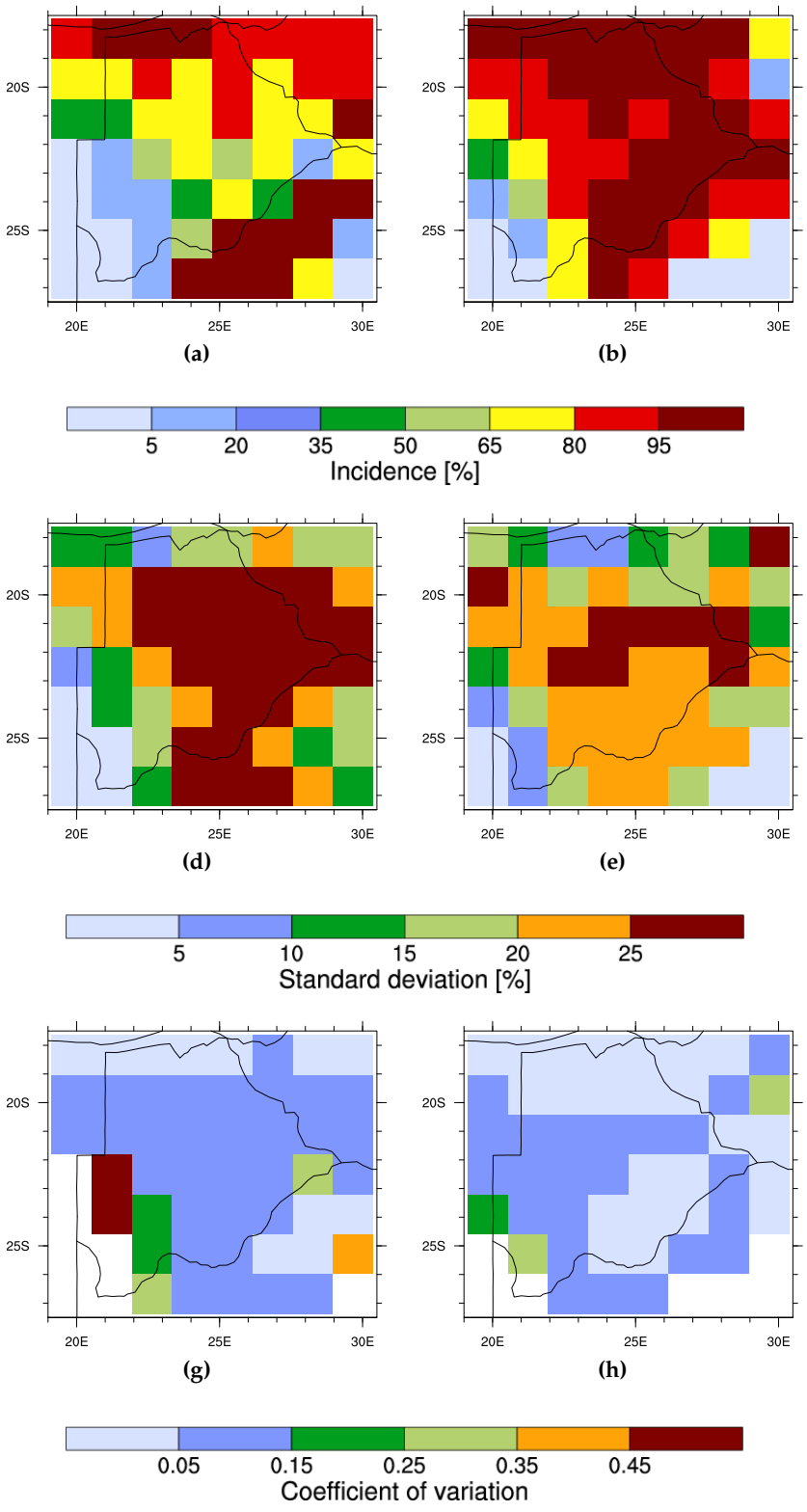


Figure 7.5: ERA-Interim (left column) and System 4 (right column) Botswana March-May malaria incidence climate (a, b), variance (c, d) and coefficient of variation (e, f), as simulated by the LMM.

which show that the ERA-Interim simulated malaria incidence is initially concentrated in the south east, then spreading to the north of the region, before abating in the south. This is similar for System 4, except for an over simulation in the north in February. This is the same area in which System 4 over predicts the rainfall in December and January, thus the wet bias is likely the cause of the bias in incidence.

Analysis of March-May incidence is shown in figure 7.5. The mean is generally correct in pattern though is too high in the centre of Botswana for System 4. The standard deviation is highest in the centre of the country, which is generally well simulated in System 4, though the magnitude of the variation is underestimated. The coefficient of variation for ERA-Interim is highest in the south west of the region and in the south east, which is similar for System 4.

Tier-2 ROC AUC maps for Botswana are shown in figure 7.6. The score is highest for forecasts issued in December, for lower tercile events. The significance scores for this prediction lie above significance for nearly the whole country, which is a region where the FMA coefficient of variation in incidence is high (figure 7.5). This suggests that a useful forecast of low malaria could potentially be made in this region. Scores for upper tercile events are also high, but not significant for most of the region. At longer lead times the skill is lower, though there is still a reasonably large area with significant ROC AUC for lower tercile events. The skill of upper tercile forecasts is lower. ROC AUC maps for bias-corrected System 4 are shown in figures 7.6d to 7.6f . The score here is similar to the raw non-bias corrected forecast, suggesting that the source of error here is not from the temperature bias.

These results are consistent with table 7.2, where tier-2 skill is higher for lower than upper tercile events, with scores higher in December. However it shows the uncertainty related to tier-2 validation: scores are significant when validated against real malaria data for November forecasts, whilst tier-2 shows low skill for these start dates. Therefore poor performance at tier-2 does not necessarily mean the system is not predicting malaria well. Conversely good tier-2 skill at does not necessarily mean real-world skill for malaria prediction. There is necessarily then less confidence in model performance when only tier-2 validation is possible, highlighting the importance for model validation of long, good quality observational data records.

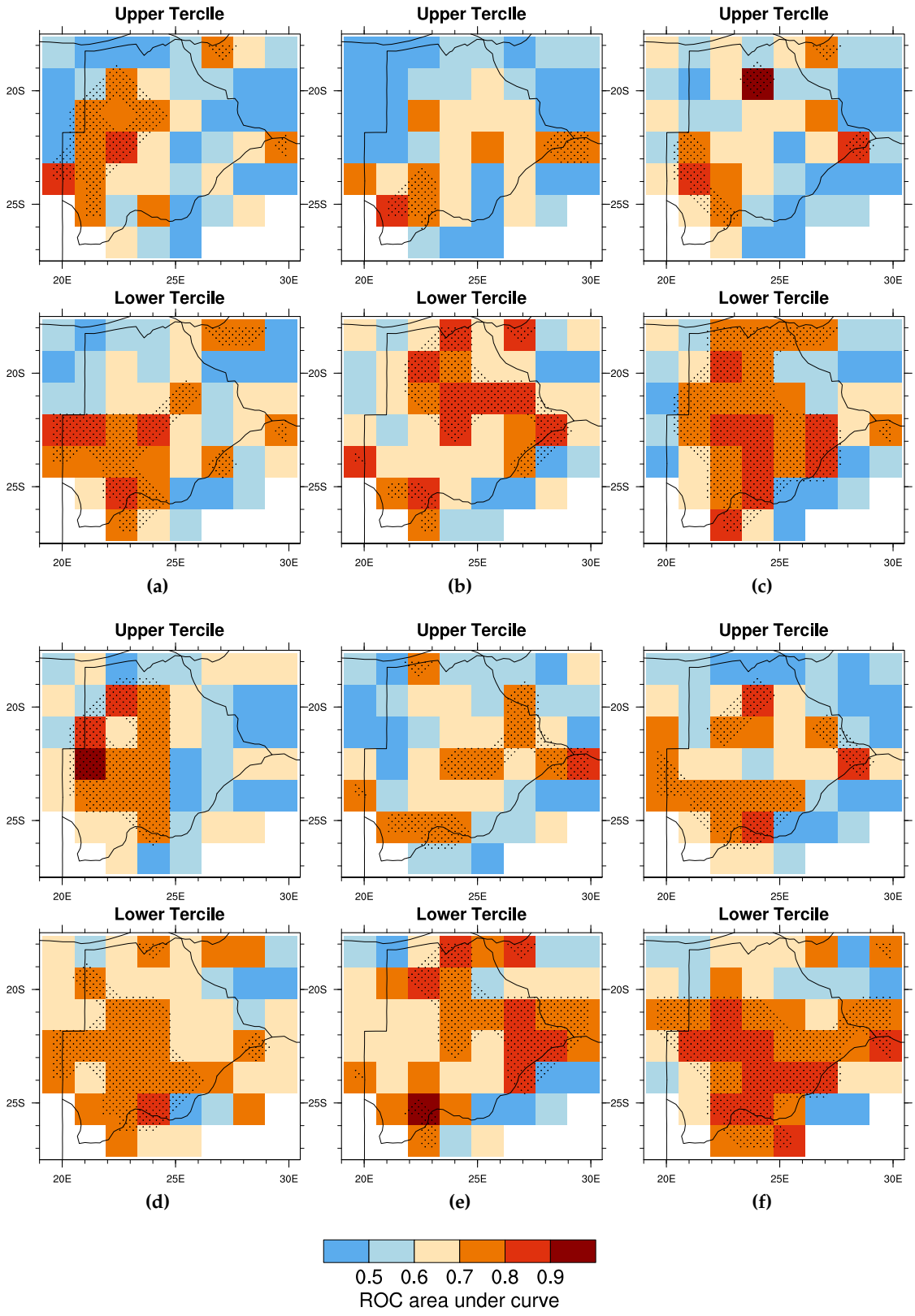


Figure 7.6: ROC area under curve for March-May malaria incidence for Botswana, LMM driven by System 4 (using LMM driven by ERA-Interim as a reference). The top row indicates forecasts made without the temperature bias correction, the bottom forecasts made with it. Forecasts shown are issued in October (a, d), November (b, e) and December (c, f), to compare with tier-3 validation, shown in table 7.2.

Sahel

The mean seasonal cycle of rainfall and temperature for the Sahel are shown in figure 7.7, for the ERA-Interim reanalysis and as simulated by System 4. ERA-I and System 4 exhibit the same general shape, with a rainfall peak centred in August with roughly the same magnitude. The rain reduces slightly more after the peak in System 4 (in September), whilst outside of June to October rainfall amounts are low. The shape of the temperature seasonality of System 4 is generally correct, however it has a moderate cold bias (around 2°C).

Maps of ERA-Interim and System 4 precipitation climatology are shown in appendix D, in figures D.6 and D.7. The spatial pattern and seasonal variation is well simulated, with a correct reproduction of the northward propagation of the Inter-Tropical Convergence Zone. The map of the temperature bias is shown in figure D.8, which in agreement with figure 7.7d is cold and over 2°C for most of the region. The bias is strongest during the boreal winter, and in the north of the region.

The simulated malaria incidence climate is shown in figure 7.7e for ERA-Interim and 7.7f for System 4. The seasonal cycle is very well simulated, in magnitude and in shape. The highest malaria incidence occurs in SON (justifying the choice of SON as a target season for the area). Spatial maps of incidence as driven by ERA-Interim and by System 4 are shown in figures D.9 and D.10. The spatial pattern is well simulated, with the epidemic fringe separating regions of constant transmission to the south and zero transmission to the north clearly defined.

Focusing on the climatology of SON, the mean, standard deviation and COV is shown in figure 7.8. COV for both ERA-Interim and System 4 is highest at a similar latitude; though for System 4 the band is at a slightly higher latitude in the west and lower in the east. This depicts the epidemic fringe in the model world, where there is significant variability in malaria incidence between years: some years experience large epidemics whilst for others the disease burden is low (Grover-Kopec et al., 2005). It is in regions like this where skilful malaria forecasts can be the most use. Compared to those in endemic malarial zones, people in low transmission areas have low immunity, increasing the impact of epidemics when the climate conditions are right.

Tier-2 ROC AUC maps are shown in figure 7.9, for May, June and July forecasts. Scores are below significance nearly everywhere. After bias correction of temperature, scores do not increase significantly. However for July forecasts there are a few grid points above significance aligning with the epidemic fringe in the west, around the border between Mali and Mauritania. These gridpoints have a high coefficient of variation (as shown in figure 7.8g), suggesting that good forecasts here are likely to be useful.

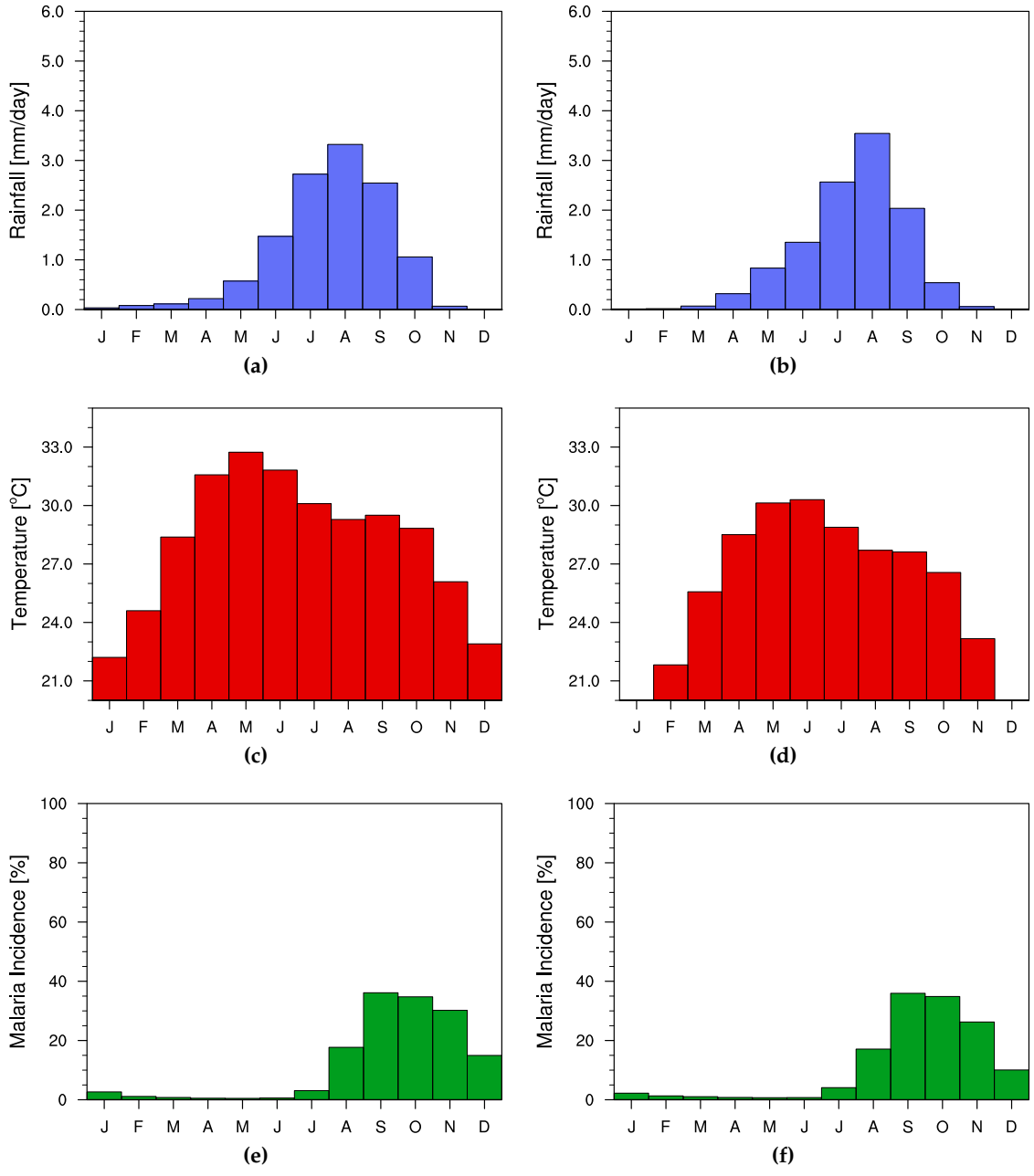


Figure 7.7: Sahel precipitation (a, b) temperature (c, d) and LMM-simulated incidence (e, f) seasonal cycle, according to ERA-Interim (left column) and System 4 (right column).

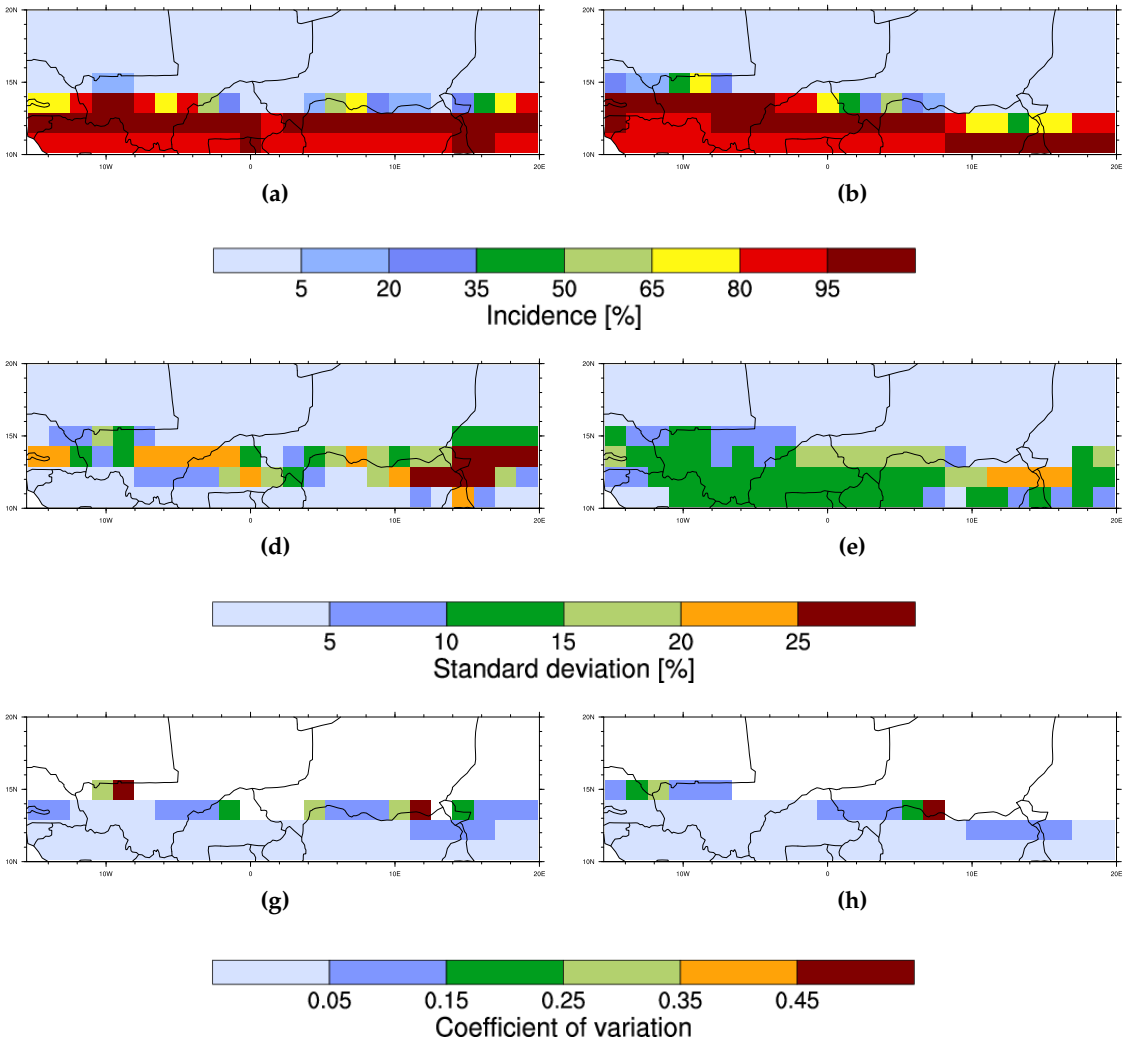


Figure 7.8: ERA-Interim (left column) and System 4 (right column) Sahel September-November malaria incidence climate (a, b), variance (c, d) and COV (e, f), as simulated by the LMM. Areas below 1% SON incidence are masked in COV plots.

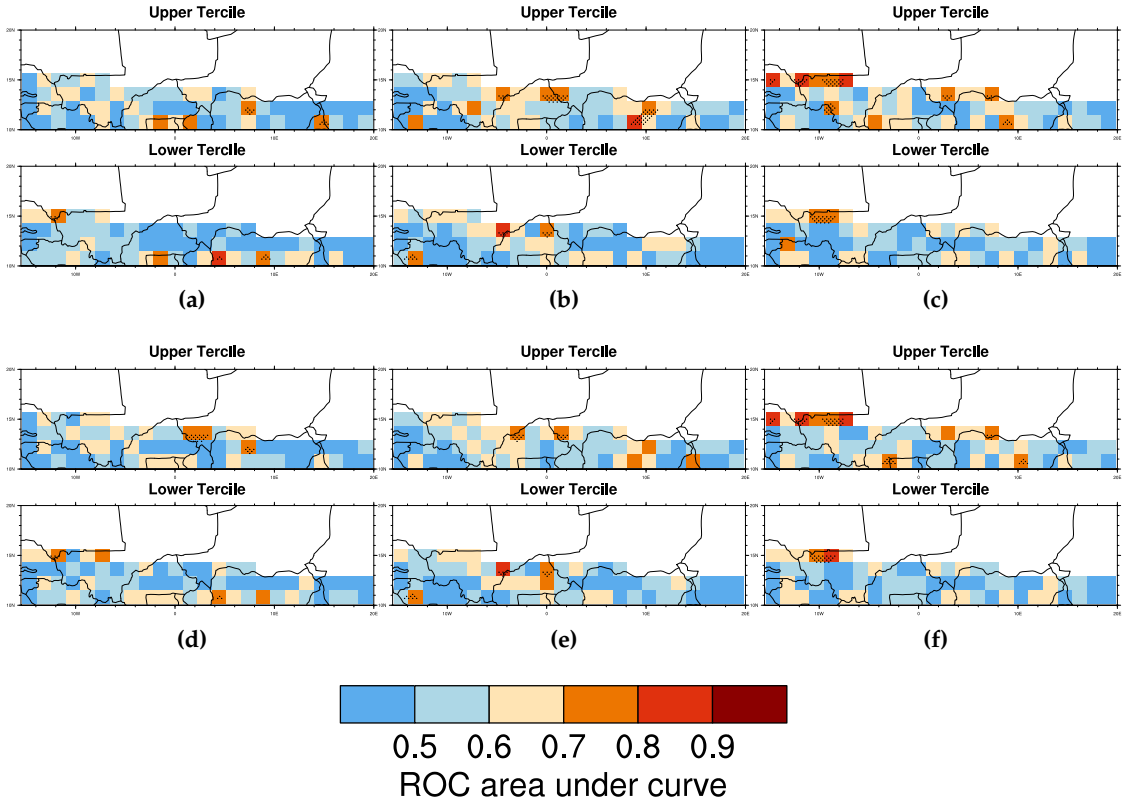


Figure 7.9: ROC AUC for September to November simulated malaria incidence for the Sahel for LMM driven by System 4 (using LMM driven by ERA-Interim as a reference). The top row are forecasts made without the temperature bias correction, the bottom forecasts made with it. Forecasts shown are issued in May (a, d), June (b, e) and July (c, f), with areas with less than 1% SON incidence masked.

Gulf of Guinea

The seasonal cycles of rainfall and precipitation for the Gulf of Guinea are shown in figure 7.10. There is a slight dry bias, though the seasonal cycle is well reproduced by System 4. There is also a cold bias, though the yearly cycle of temperature is well simulated, with maximum temperature falling in March and a minimum in August. Precipitation climatology maps are shown in appendix D, in figure D.11 for ERA-Interim and D.12 for System 4. The spatial pattern of precipitation is generally correct, with the onset and northward propagation of the rains occurring at the correct time of the year, between February and May. The climatology map of the temperature biases is shown in figure D.13. The bias is cold throughout the year and coldest outside of the rainy season; during the rainy season it is under 1°C. The spatial pattern of the bias is not consistent through the year, for instance in October and November it is largest near the coast, whilst in December January and February it is largest in the north of the region.

Malaria seasonal cycles are shown in figure 7.10e for ERA-Interim and figure 7.10f for System 4. Simulated malaria is present throughout the year. For LMM driven by ERA-Interim it occurs in two phases; January until May where incidence lies around 60%, whilst June until December has higher incidence around 80%. For System 4 the simulated malaria incidence is too low between January and May (below 40%). For the rest of the year it is higher, and roughly at around 80% (except for a peak of nearly 100% incidence simulated in July).

Maps of simulated malaria throughout the year are shown in figure D.14 for System 4 and D.15 for ERA-Interim. System 4 simulates low incidence from January until May, except for at grid points close to the coast where incidence is high all year round. During this period the simulated ERA-Interim incidence extends further inland. For the rest of the year, simulated incidence over 80% for both System 4 and ERA-Interim, with low spatial variability.

Results focusing on the SON season are shown in figure 7.11. ERA-Interim and System 4 have a similar mean malaria incidence, standard deviation and coefficient of variation. The results suggest that there is a very low variability in SON incidence over the region. High incidence and low standard deviation create a very low coefficient of variation, denoting stable transmission. This is an endemic profile which has been shown in other modelling and observational malaria studies (eg. Snow et al., 1999).

ROC AUC for SON for the Gulf of Guinea are shown in figure 7.12 for raw and bias corrected forecasts. There is little difference in the magnitude of the score; for raw and

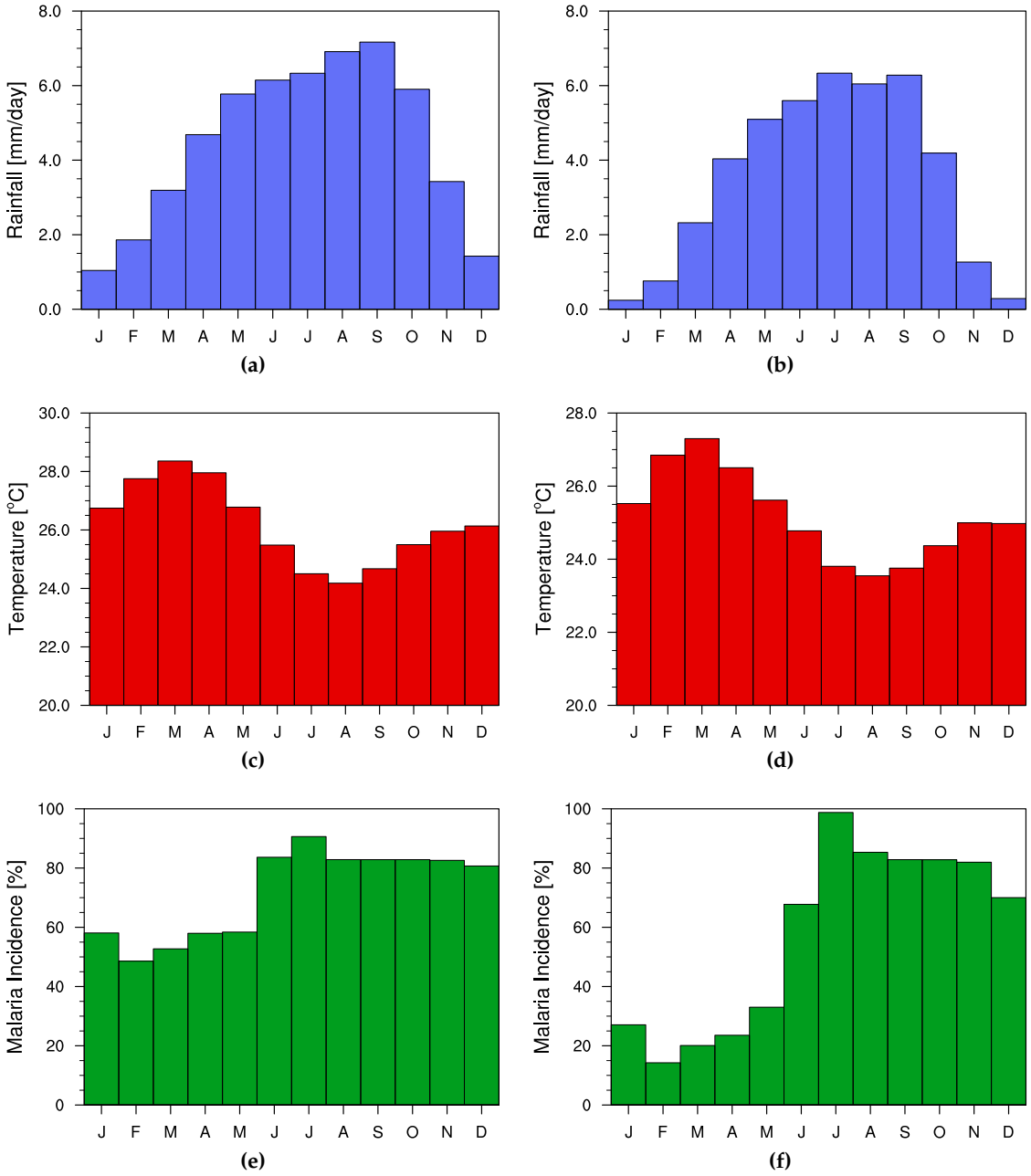


Figure 7.10: Gulf of Guinea precipitation (a, b) temperature (c, d) and LMM-simulated incidence (e, f) seasonal cycle, according to ERA-Interim (left column) and System 4 (right column).

.

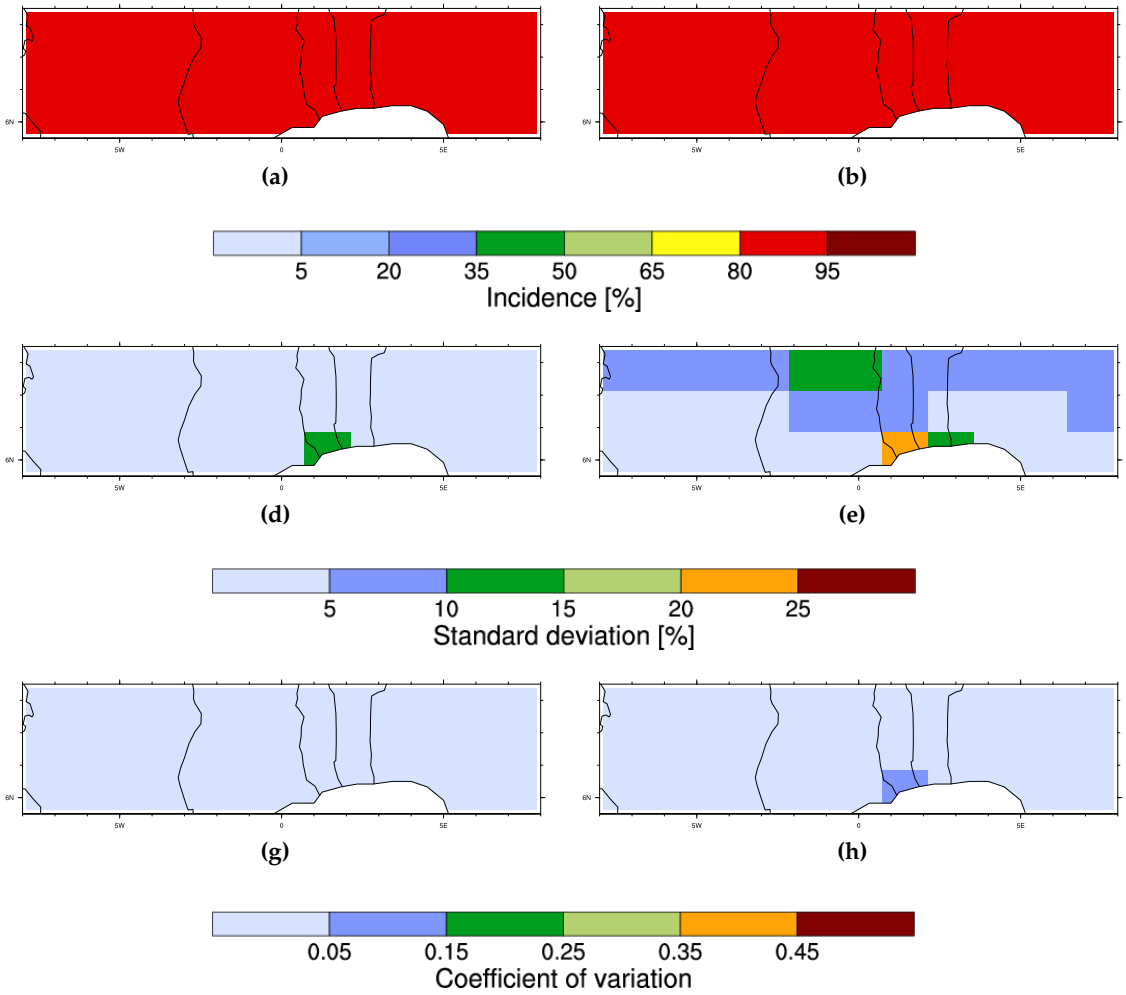


Figure 7.11: ERA-Interim (left column) and System 4 (right column) Gulf of Guinea September to November malaria incidence climate (a, b), variance (c, d) and COV (e, f), as simulated by the LMM. Areas below 1% SON incidence are masked in COV plots.

corrected forecasts, both upper and lower tercile forecasts at all lead times have ROC AUC below significance.

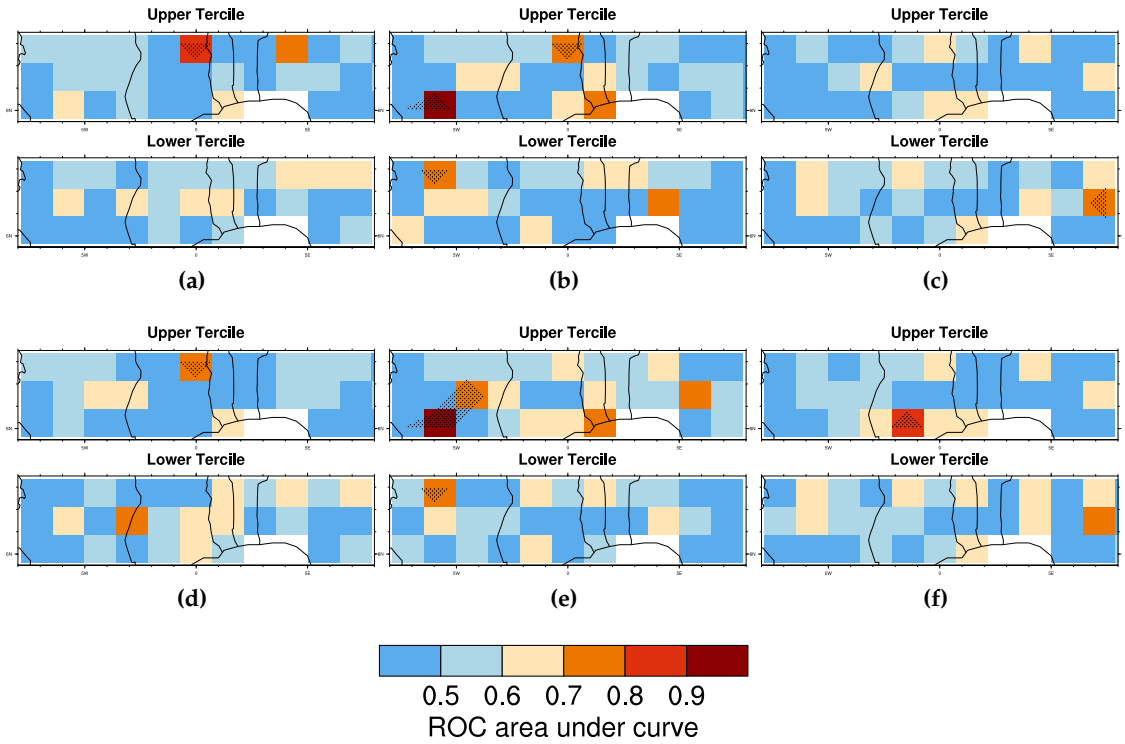


Figure 7.12: ROC AUC for September to November malaria incidence for Gulf of Guinea, LMM driven by System 4 (using LMM driven by ERA-Interim as a reference). The top row are forecasts made without the temperature bias correction, the bottom forecasts made with it. Forecasts shown are issued in May (a, d), June (b, e) and July (c, f), with areas with less than 1% SON incidence masked.

Malawi

Figure 7.13 shows the mean seasonal cycle of rainfall and temperature over Malawi. System 4 is able to realistically capture the observed mean seasonal cycle, however rainfall is generally overestimated and temperatures are consistently too cold, by about 1°C.

Maps showing the precipitation climatology over the country are displayed in appendix D, for ERA-Interim in figure D.16 and for System 4 in figure D.17. The rainfall season according to ERA-Interim falls in November to March, and this is well simulated by System 4. Despite this there is a wet bias throughout the rainy season. From April to October the amount of rainfall is negligible in both the model and the reanalysis. The spatial pattern of rainfall is such that there is a local minima across two gridpoints in the centre of the north part of the domain, over the northern edge of lake Malawi. This feature is not captured by System 4, which may be related to the representation of topography in the model, as the north west of the country is mountainous with the altitude dropping toward lake Malawi.

The mean bias in temperature is plotted in figure D.18. The cold bias is generally smaller during the dry period of the year and larger during the months with significant rainfall. This is most likely due to the wet bias; consistently higher rainfall will reduce the average temperature. Furthermore this wet and cool bias is generally stronger over the western part of the domain, for high altitude regions, suggesting model problems in reproducing rainfall over the mountains.

The malaria season simulated by ERA-Interim over Malawi ranges from February to July (figure 7.13e); during this period the highest malaria incidence (>60%) is from March to May. The mean seasonal cycle of malaria incidence simulated by System 4 is relatively realistic with respect to ERA-Interim; a large incidence is simulated from February to July, with a peak occurring in March (figure 7.13f). However, the mean System 4 malaria incidence is generally underestimated, except in February when it is highly overestimated. This is relatively consistent with the differences rainfall and temperature seen between System 4 and ERA-Interim over this region; slightly cooler temperatures decreasing the simulated malaria incidence and the wetter conditions occurring two months before the incidence maximum (in December to January) increasing the magnitude of the malaria incidence peak.

Maps of the seasonal cycle of incidence are shown in figure D.19 for ERA-Interim and in figure D.20 for System 4. Malaria is present in the south east of the region only in January and February, before spreading northward and covering the whole region in April and May. It then reduces in the north west and finally drops to around zero by

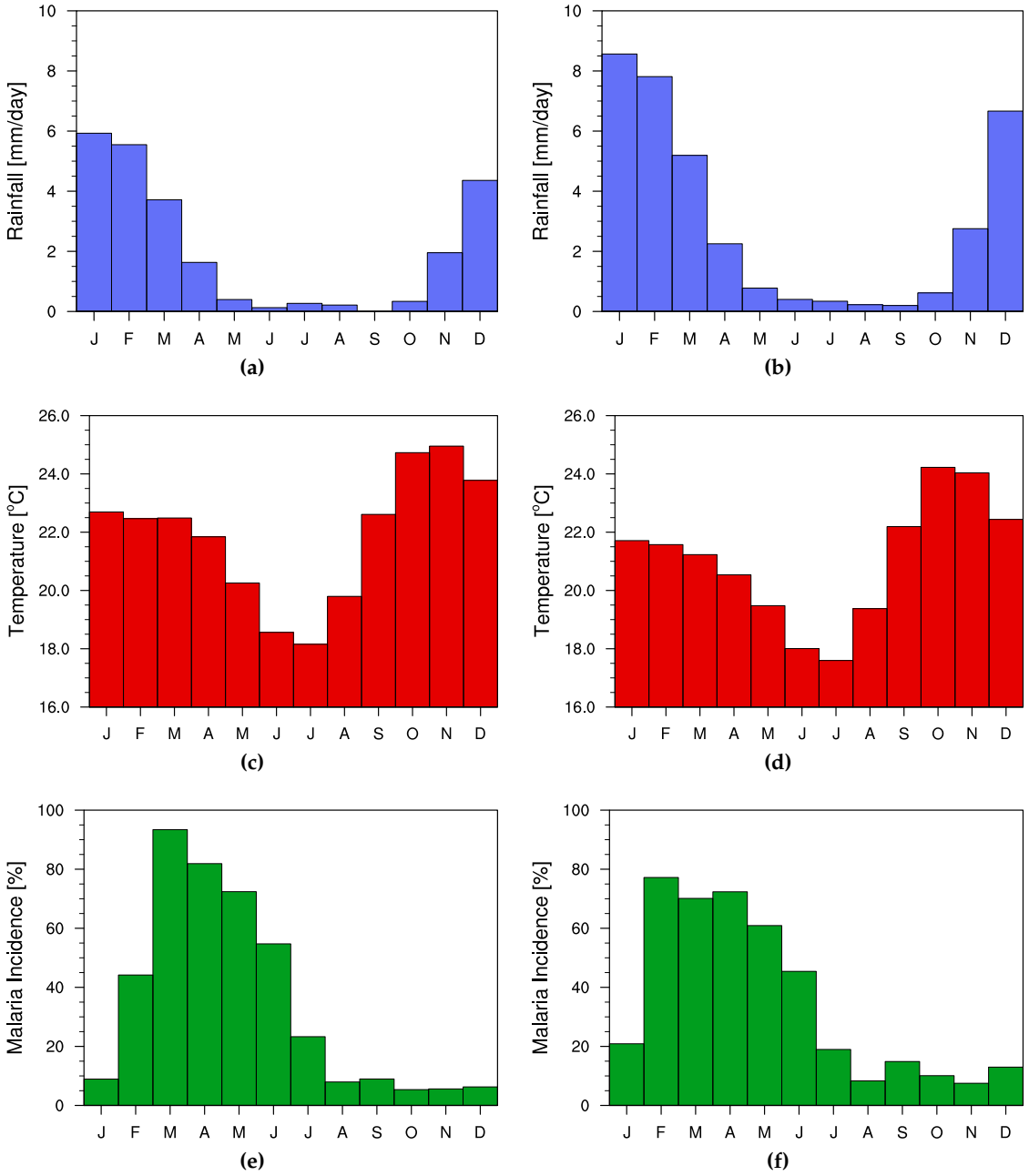


Figure 7.13: Malawi precipitation (a, b) temperature (c, d) and LMM-simulated incidence (e, f) seasonal cycle, according to ERA-Interim (left column) and System 4 (right column).

.

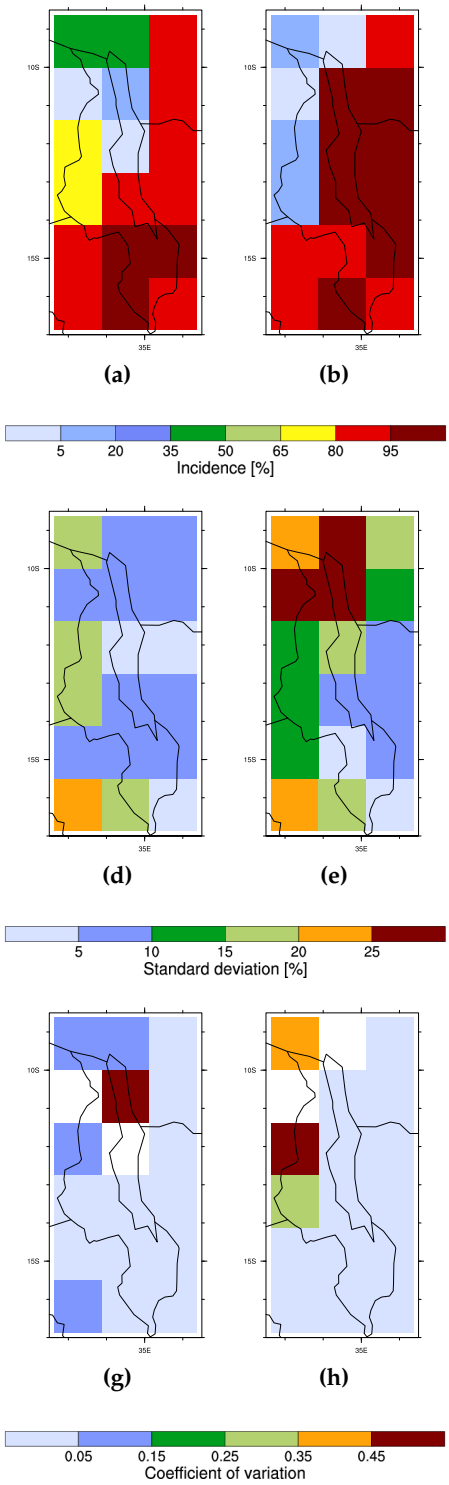


Figure 7.14: ERA-Interim (left column) and System 4 (right column) Malawi March to May malaria incidence climate (a, b), variance (c, d) and COV (e, f), as simulated by the LMM. Areas below 1% SON incidence are masked in COV plots.

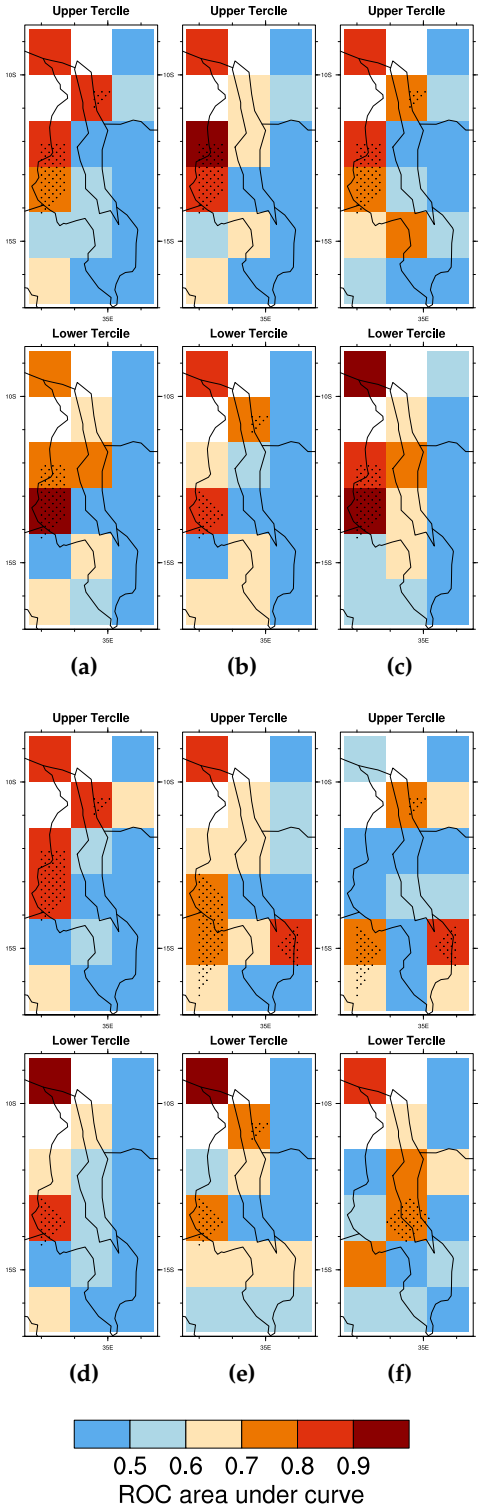


Figure 7.15: ROC area under curve for March to May malaria incidence for Malawi, LMM driven by System 4 (using LMM driven by ERA-Interim as a reference). The top row are forecasts made without the temperature bias correction, the bottom forecasts made with it. Forecasts shown are issued in October (a, d), November (b, e) and December (c, f).

August. System 4 simulates this well, except for February where the incidence moves too far north too quickly, and in March to June, where it is too low in the north west.

Mean malaria incidence for MAM is plotted in figure 7.14. System 4 over-predicts incidence in the central-eastern side of the domain with respect to ERA-Interim, whilst the meridional gradient in incidence (with higher malaria from south to north) is relatively well reproduced by System 4 over the western half of the domain. The largest COV is simulated by both ERA-Interim and System 4 lying over the north west of the region, indicating an epidemic profile for the region here.

Figure 7.15 shows tier-2 ROC AUC for Malawi, for forecasts issued in October, November and December. Generally, the skill in malaria incidence increases as a function of lead time over Malawi, and skill is slightly higher for upper tercile than for lower tercile events, particularly at longer lead times. ROC AUC is above significance over the north west of the region, for both upper and lower tercile forecasts, showing potential forecasting ability for malaria events over those areas (in a tier-2 validation context). The region where the ROC area is above significance covers the north-western part of the domain in an L-shape. This area of large skill is an area which exhibits a large COV (figure 7.14). The skill of System 4 in forecasting malaria events is thus larger over the epidemic areas of the Malawi domain, suggesting useful forecasts can potentially be made here.

There is no overall increase in ROC area with bias correction, and whilst some forecasts are slightly improved with the correction (e.g. lower tercile forecasts issued in November, figure 7.15e), others are significantly degraded (e.g. forecasts issued December, figure 7.15f).

7.2.3 Quantifying uncertainty in the prediction of climate-driven disease risk

System 4 has been validated at tier-3 level for Botswana and at tier-2 for other regions. The central question of the thesis still remains: how can uncertainty climate-driven disease risk forecasts be quantified? Posing the question an alternative way: given validation results for a hindcast period, what is the confidence in a new forecast? This theoretical results section briefly considers this question.

Disease risk can be defined as an upper tercile incidence event (conversely, a lack of risk is defined as a lower tercile event). To relate uncertainty in predicting this risk to model validation, some general principles can be stated. Firstly, the historical success at predicting an event should be proportional to confidence in a new forecast. i.e. if every time our forecast system predicted 'event', an event did indeed occur, then our

confidence in a new forecast of 'event' is high. On the other hand, if of all the times the system forecast predicted 'event', the event rarely occurred, our confidence in a new forecast is low. Quantifying this we can define the 'hit fraction', H_f, i.e. the number of hits divided by the sum of hits and false alarms;

$$H_f = \frac{a}{a+b}\,, \tag{7.1}$$

where a and b are the number of hits and number of false alarms respectively defined previously, in table 3.2. Note that this is different to the hit rate H, which is conditioned on observed rather than forecast events. H_f is also equal to one minus the false alarm rate.

A relative measure of confidence, C, can then begin to be defined, as

$$C \propto H_f \times 100\%\,, \tag{7.2}$$

where scaling is applied so C ranges from 0 % to 100%, percentages being a familiar way to measure confidence.

A second principle can be stated: the confidence that hindcast validation of a forecast system reflects its underlying quality should be proportional to the size of the validation sample, N, where

$$N = a + b\,. \tag{7.3}$$

That is, if the system only predicted an event once in the hindcast period, then confidence in a new forecast is low, even if that one event validated. A larger number of forecast events in the hindcast sample gives more confidence that we understand the behaviour of the model. For example, we have likely more confidence we understand the behaviour or the system when $N = 10$ and $H_f = 0.6$, compared to when $N = 2$ and $H_f = 1$.

How might this be introduced to equation 7.2? Confidence should be proportional to a component dependent on N, which equals zero when $N = 0$ and approaches one as $N \to \infty$, for example;

$$C \propto \left(1 - e^{-N/R}\right)\,. \tag{7.4}$$

Here the term R has been introduced as a 'risk aversion' factor to allow for adjustment of confidence on a case-by-case basis. This is necessary as different situations require

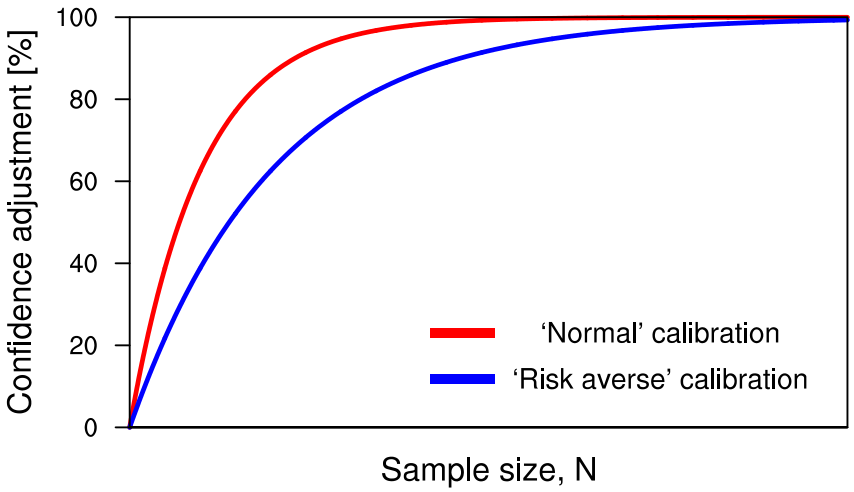


Figure 7.16: Adjustment to forecast confidence based on the number of forecast events in a hindcast sample. The red (blue) curve represents a smaller (larger) value of the risk aversion factor R; correspondingly an issued statement of confidence is be adjusted further downwards with a larger risk aversion.

calibration of confidence; if the situation involves human life and a bad forecast could lead to loss of trust in the forecast provider then perhaps one may want to select a higher value of R. Using higher values of risk aversion correspond to a more cautious assessment of confidence. This adjustment related to sample size is shown in figure 7.16.

This confidence adjustment reduces C below 100% which is rational - on scientific principles a model forecast should never come with 100% confidence, even if it has validated perfectly over a large hindcast sample.

Combining equations 7.2 and 7.4 then gives a full equation for confidence in forecasts of an event;

$$C = H_f(1 - e^{-N/R}) \times 100\% \,. \tag{7.5}$$

Note that this confidence measure will be different for forecasts of non events over the same hindcast period, as a 'correct rejection fraction' rather than 'hit fraction' should be used, and the sample size will be different. The method however remains the same.

Finally we consider a further adjustment of confidence in the case of tier-2 validation. Qualitatively it can be stated that we have more confidence in a climate-driven disease forecast with good skill at tier-3 level (against real observations of disease incidence) than we do if it shows the same level of skill at tier-2 level (against reanalysis-driven model forecasts). How then can this difference be quantified?

A term to account for the confidence that reanalysis-driven model represents real disease data should be introduced to equation 7.5. If the reanalysis driven model output is close to disease data then we have more confidence in tier-2 validation; if there is no correlation between the timeseries then we have no confidence that tier-2 validation is meaningful. A correlation coefficient, for example Pearson product-moment correlation coefficient r, is then appropriate, so we can define the relative confidence in tier-2 validation as

$$C_{t2} = rH_f(1 - e^{-N/R}) \times 100\% \,, \tag{7.6}$$

where confidence in tier-2 validation will be zero if disease model driven reanalysis has no correlation to real disease data, and will be equal to tier-3 validation confidence if it exactly matches.

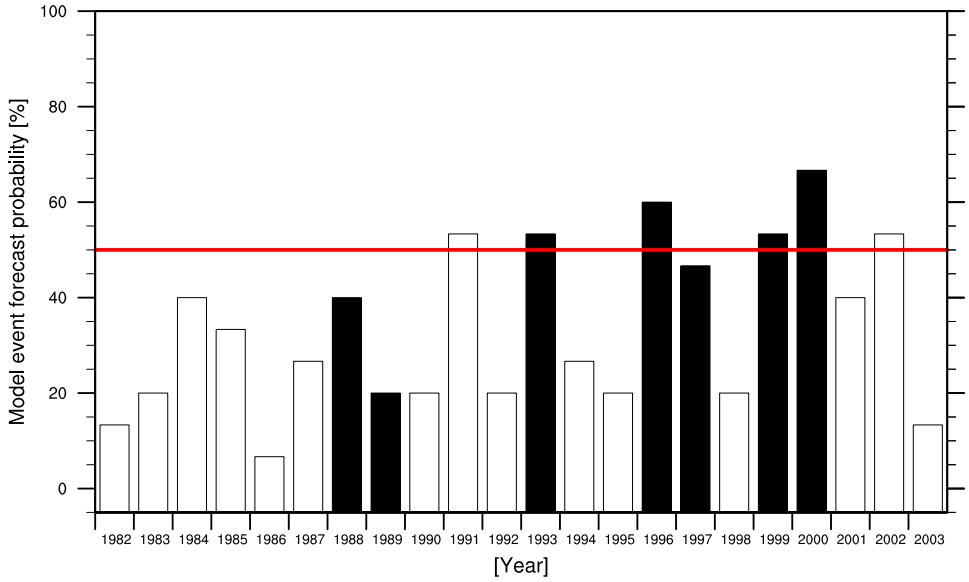
To demonstrate an application of this method it is here applied to LMM upper tercile incidence forecasts in a tier-3 context over Botswana and at a tier-2 context over Malawi. Forecast probabilities are shown in figure 7.17, where hits and misses are defined using the decision thresholds indicated by the red line. The decision threshold here is chosen as 50%, that is, a forecast of 'event' is issued when at least half the ensemble members lie in the upper tercile category. An optimisation of the decision threshold is possible by analysing economic value curves, though this is not carried out here.

Looking first at tier-3 validation over Botswana (figure 7.17a), H_f can be calculated as $4/(4+2) = 0.67$. If R is taken as 1 and the sample size is 6, this gives a confidence in forecasts of 67%[3].

For Malawi, H_f is $7/(7+3) = 0.7$ and with $R = 1$ and $N = 11$ this gives confidence as 70%. However the tier-2 correlation correction in equation 7.6 applies. This cannot be calculated specifically for Malawi, but instead it can be estimated from where there is tier-3 data, over Botswana[4] The correlation between LMM-ERA-Interim and the BMI is 0.65, reducing the confidence in predictions of upper tercile events such that $C = 46\%$. This is much lower than the confidence in upper tercile forecast of Botswana, despite

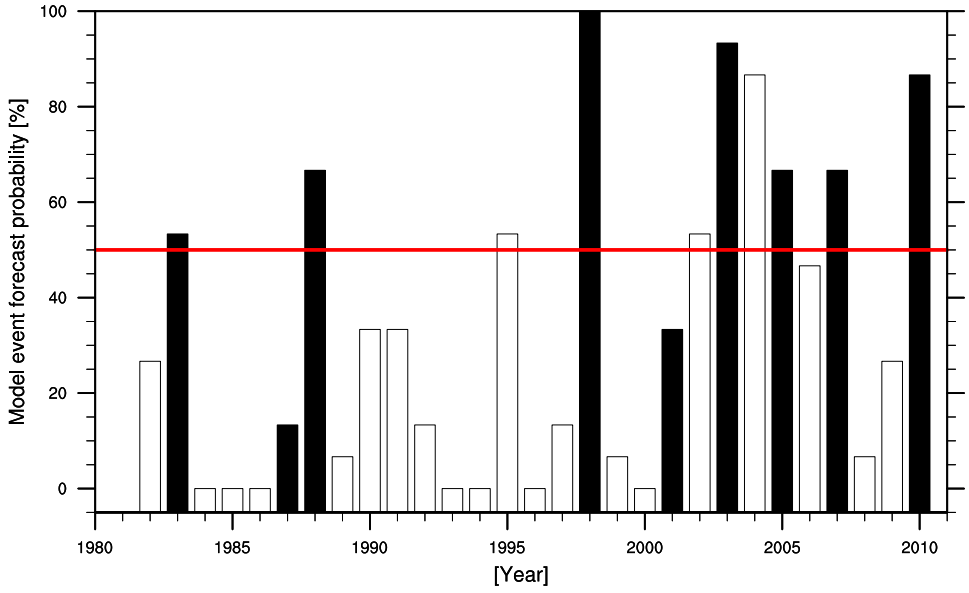
[3]The reduction due to sample size in this case does not produce a significant correction for $N > 2$. It may be that a larger correction is needed and it would be possible to do this by calibrating R, the risk-aversion factor. How exactly this parameter might be calibrated is an open question, but could be done by calculating the half-life, i.e. how many forecast events are needed such that we are more than 50% sure that our validation is capturing the behaviour of the system.

[4]Note that this an extrapolation: this correlation coefficient is calculated over Botswana and is likely to be different for Malawi. In the absence of Malawi malaria data however it is arguably the best measure of the mismatch between ERA-Interim and real malaria incidence. Furthermore, if one had observed malaria data for Malawi, measuring the discrepancy between LMM driven by reanalysis and observed malaria cases would be unnecessary as one would be able to validate the LMM at a tier-3 level.



4 hits
3 misses
2 false alarms
13 correct rejections
ROC area under curve = 0.86

(a)



7 hits
2 misses
3 false alarms
17 correct rejections
ROC area under curve = 0.89

(b)

Figure 7.17: Forecast probability barcharts for upper tercile (a) January to May Botswana incidence forecasts issued November (tier-3, vs. BMI) and (b) March to May Malawi incidence forecasts issued December (tier-2, vs. ERA-I).

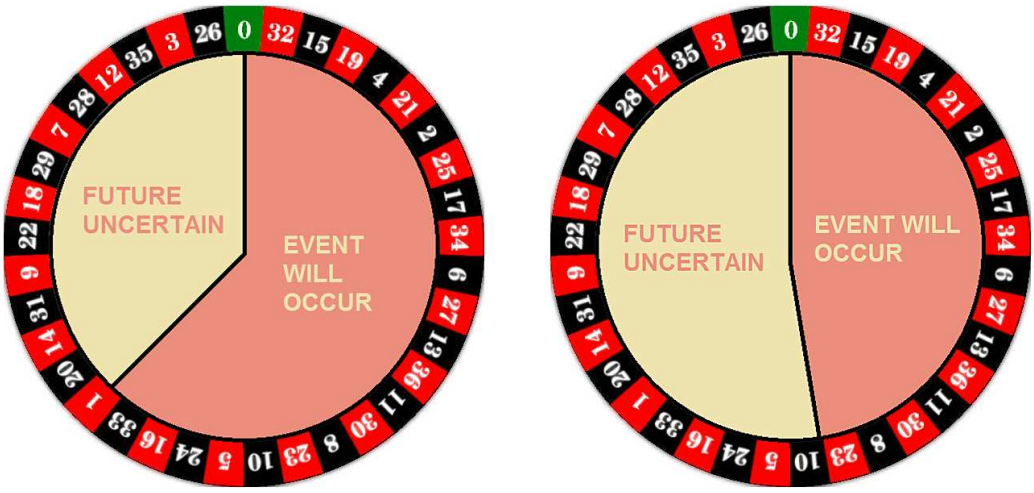


Figure 7.18: Communicating uncertainty associated with a forecast of 'event': on the left the wheel corresponding to a Botswana forecast (confidence 67%) and on the right a forecast associated with a Malawi event forecast (confidence 46%).

a similar hit fraction; however is representative of the uncertainty in a tier-2 validation context.

Note that these confidence percentages are not equivalent to our estimated probability of the event. Rather, the method works on a principle that a 'deterministic' forecast of 'event is issued (based of course on a probabilistic ensemble). A confidence of 67% does not mean that 'when we predict an event 67% of the time the event will happen and 33% it will not', it is instead like saying 'when we predict 'event' we believe that 67% of the time it will happen and 33% of the time anything could happen'. Inspired by previous efforts (Hagedorn and Smith, 2009) at communicating the value of probabilistic forecasts (and gambling), this confidence for the forecasts considered above is illustrated in figure 7.18; for Botswana with a roulette wheel on which 67% of the divisions is written '**EVENT WILL OCCUR**', and the remaining 33% saying '**FUTURE UNCERTAIN**'. The wheel for Malawi would then have only 46% of divisions saying '**EVENT WILL OCCUR**'.

Thus a decision maker could be told that if our modelling system predicts event, it is like spinning the corresponding wheel. THis intuitively builds in the communication of uncertainty into the communication of the forecast, giving the decision-maker a realistic honest understanding of our confidence in the forecast system.

7.3 Discussion

System 4 has been validated against malaria data over Botswana and skill has been demonstrated for forecasts issued in November. These forecasts have value according to a model of cost/loss and are potentially ready for use by relevant humanitarian organisations working in the area. Previous work validating the LMM against the BMI, driven by DEMETER forecasts (initialised in November) gave ROC AUC for upper and lower tercile incidence forecasts as 0.67 and 0.84 (Jones and Morse, 2010). Comparing these values with the corresponding forecasts from System 4 (0.85 and 0.80 respectively) shows that System 4 offers significantly improved upper tercile forecasts, with no improvement for lower tercile events[5].

For regions without available malaria data, tier-2 validation has been carried out. For the Sahel potential skill in predicting malaria was shown for forecasts issued in July, around the border between Mali and Mauritania. This area shows an epidemic profile and skilful prediction here is consistent with previous work with the LMM (Jones and Morse, 2012). Defining a smaller subregion around the fringe where ROC AUC is high is then a next step toward the production of useful forecasts here.

For Malawi, MAM incidence forecasts issued in December are also skilful in the north west for both upper and lower tercile events, again in a region where the coefficient of variation is high. Analysis of the drivers also showed that System 4 has a wet and cool bias in the north west, likely to be due to poor model representation of topography (related to the spatial resolution of the model). Therefore forecasts here may be further improved by using precipitation bias corrected input (the potential for this is discussed below).

No skill was found for SON incidence in the Gulf of Guinea. Here the coefficient of variation is low, suggesting there even if it were possible to make skilful forecasts of seasonal average malaria they would not be useful. Low skill in this region is consistent with former results using the LMM (Jones and Morse, 2012; Jones and Morse, 2010), which suggest that due to the lack of an immunity component LMM forecasts are not skilful in regions where there is year-round transmission[6]. However it may be that malaria early warnings could be useful, for instance if it were possible to predict the onset of the main malaria season. Further work, introducing immunity to the LMM and

[5]Note that the sample size in the present analysis is slightly larger than previously used (since as the DEMETER hindcast period ends in 2001). The ROC AUC was recalculated using the same index of years, giving scores for upper and lower tercile forecasts as 0.78 and 0.95. Therefore the conclusion can be made with confidence that System 4 makes better malaria forecasts over Botswana than DEMETER.

[6]A caveat should be placed on results here: since validation is at tier-2 level then any problems the LMM has with endemic areas will affect the simulated malaria in the target reanalysis-driven runs, reducing the realism and significance of any result at tier-2.

informing potential forecast targets with local knowledge could potentially uncover skill in this area.

An important caveat associated with tier-2 results relates to the ERA-Interim reanalysis. This dataset was used as a reference as it provides gridded datasets for the time period that fits the ECMWF System 4 hindcasts. However, unknown biases may exist in these reanalysis, that is, there is no way to verify reanalysis in places and at times where no observations exist. In fact over Malawi the monthly rainfall cycle simulated by System 4 is closer to the CRUTS2.1 dataset (based on observations) than the ERA-Interim reanalysis is (QWeCI Project, 2011). This adds an unknown element of uncertainty into forecasts; if a target is not reality there is uncertainty that a forecast with skill can actually make predictions of reality, and that a forecast without skill cannot. To explore this uncertainty, the analysis contained in this chapter should be repeated with other daily reanalysis datasets, testing the robustness of skilful results.

System 4 also has biases in temperature and precipitation. It is unknown to what level these errors propagate to malaria forecasts, and is likely to be significant in areas where the average temperature is close to the sporogonic threshold (described in section 7.1.1). A simple temperature bias correction method was applied and is shown to improve forecast quality for Botswana, though had limited (and occasionally negative) effects elsewhere. The same methodology for bias correction cannot be applied to precipitation biases; the ubiquity of days of zero rain would lead to negative values of rainfall. A simple precipitation bias correction applied previously to LMM driving data did not improve skill (Jones, 2007) and it is likely that a more complex bias correction of precipitation is necessary in order to extract all potential skill from the climate forecasting system. On this point, calibrated System 4 hindcasts have been produced at the ECMWF, using a bias-correction technique based on empirical orthogonal functions. Malaria forecasts with the LMM are likely to be improved with these recently completed calibrated hindcasts. A comparison of LMM forecasts driven by uncalibrated and calibrated System 4 hindcasts should be a priority step for the continuation of this work.

The link between skill at tier-1 and tier-2/3 is not clear. *A priori* one may think that higher skill at predicting the driving climate will result in better malaria forecast. For instance this is the case for Botswana, where November climate forecasts have higher value for upper tercile temperature and precipitation than those made in October or December (see table 6.2), correspondingly malaria forecasts over Botswana are best in November. However, when considering the area of the Sahel which demonstrates skill around the epidemic fringe at tier-2 level for forecasts issued in July (figure 7.9c), this area has the highest ROC AUC overall at tier-1 level (temperature and precipitation, upper and lower

tercile) not for the July but instead for the May forecast start dates. ROC AUC for July start dates is in fact below significance here for upper tercile precipitation forecasts.

The reason for poor performance at tier-2/3 despite good performance at tier-1 is not clear. One reason may be methodological, that the metrics are not valid. The three-month season targets for tier-1 validation were selected based on the maximum of the rains, i.e. the peak of the rainy season. The rationale was that the season with the highest mean and variance of rainfall drives malaria outcomes. However this may not necessarily always be the case, especially for endemic regions where the rainfall season lasts much longer than three months (e.g. in the Gulf of Guinea), or where aspects of the temperature cycle are more relevant for the disease, such as the east African highlands (Pascual et al., 2006). This may be the reason why skill is observed at tier-2/3 whilst not being present at tier-1; System 4 forecasts may contain skilful information outside of the targets defined here.

Another reason for good tier-1/poor tier-2 validation is unrealistic simulation of intra-seasonal dynamics. A climate model may forecast a three month seasonal average accurately, whilst unrealistically simulating the day-to-day variation of temperature and rainfall. For example, a forecast of a three month season with 90 millimetres accumulated rainfall could correspond to one millimetre every day, or 90 millimetres on the first day followed by 89 dry days. Since the LMM takes daily temperature and rainfall as input, errors in the day-to-day variation of the weather may interact with the non-linear nature of the model, causing unpredictable error emergence.

This is possibly one reason why malaria forecasts do not validate whilst climate forecasts do; that is, the LMM poorly simulates malaria due to to unrealistic System 4 intra-seasonal variability, despite correct seasonal averages. Conversely, tier-2/3 may also have skill due to skilful prediction System 4 skilfully predicting sub-seasonal dynamics, despite incorrectly simulating seasonal totals. If this is true it would exist at odds with seasonal predictability theory, stating that the source of predictability comes from slowly varying components of the atmosphere and ocean, anomalies which persist for months on average, affecting the average climate of a region. The mechanism by which sub-seasonal variability may be predictable on seasonal scales whilst seasonal totals are not is not clear.

Despite extensive work and validation with the LMM here and elsewhere (Ermert et al., 2011a,b; Ermert et al., 2012; Hoshen and Morse, 2004; Jones and Morse, 2012; Jones and Morse, 2010; Jones, 2007; Thomson et al., 2006) the question of the source of skillful LMM forecasts remains open. Necessarily it must be some aspect of the input temperature and precipitation timeseries which makes the difference between a skillful and an unskillful forecast, however as discussed above it is not necessarily only seasonal totals which contribute to skill; other factors are likely to have an effect. Further work looking in

detail at the mechanics of the LMM is needed if the source of predictability is to be found.

The link between seasonal average climate and malaria is a question further explored in the next chapter. The question of whether malaria forecasts based primarily on seasonal average climate information can be made is the main question of the investigation. Quantifying uncertainty from the malaria model is also addressed.

CHAPTER 8

Relating seasonal average climate to malaria risk

This chapter considers relationship between low temporal resolution climate information and malaria outcomes, as simulated by the Liverpool Malaria Model (LMM). The investigation also considers the question of how to quantify disease model uncertainty and communicate quantified uncertainty information.

8.1 Introduction

Seasonal forecasts generally advise of conditions averaged over several months. For example, the main European and U.S. climate prediction centres currently offer forecasts of three month averages; information at a higher resolution than this is not provided, that is, model predictions of sub-seasonal variability are not available. Even though climate models run on a sub-daily timestep, i.e. they do in fact make 'predictions' of daily and sub daily variation months in advance, this is not released to users.

It is well understood that seasonal predictability comes from slowly evolving modes of the climate system and is associated with a long term persistent anomaly over a region. As such there is no strong predictable signal for sub-seasonal variability, at months in advance (see section 2.1.2 for a discussion of this point). This then is the reason for the lack of sub-seasonal variability forecasts.

On the other hand, the processes which make up the malaria transmission cycles evolve on daily and sub-daily timescales. The gonotrophic and sporogonic cycles are dependent on daily temperature and the ponds forming mosquito breeding sites can grow or shrink

on a daily timescale. This is reflected in the LMM, which takes as input daily temperature and rainfall.

There exists then a timescale separation. On one hand, to predict malaria one requires a forecast of the daily evolution of temperature and precipitation; on the other, the best forecasts available from seasonal climate models are for seasonal averages, and at higher temporal resolution there is not an established source of predictability.

This separation inspires the work contained in this chapter, and the question: is it possible to relate low temporal climate information (i.e. seasonal averages) to malaria outcomes? That is, if one only has a climate forecast of the average climatic conditions of the upcoming rainy season can something useful still be said about the associated malaria? If so, what is the uncertainty?

8.2 Methodology

To explore this question, the LMM was used as representative of the malaria transmission cycle (details of the model can be found in section 7.1.1). The methodology is to drive the LMM with multiple climate time series, and to study its response. In each instance seasonal averages of the climate inputs and incidence outputs are taken. The data is then interrogated.

To drive the model, the 20th century reanalysis dataset was used. This is a reanalysis dataset produced by NOAA, covering the period 1871-2010, further details are given in section 3.1. The reason for using this dataset is the large period, which allows for a greater sample to run through the LMM[1].

Using this reanalysis introduces the assumption that the daily evolution of rainfall and temperature in the reanalysis world is representative of reality. There are two alternatives which would eliminate or reduce this assumption; one is to use station data; the other is to generate synthetic data. Whilst station data would be more realistic, the number of years available is generally limited, preventing a full exploration of LMM behaviour. Generating synthetic data instead would remove the limit on availability of input, but employing a simple generation method is unlikely to be close to reality. Perhaps with the adoption of existing methods (such as weather generators), limitless synthetic time series could be generated to more systematically explore the

[1]Note that this investigation does not require that the reanalysis is representative of 'what really happened' i.e. that the total precipitation for Liverpool on 1st April 1882 is indeed the amount suggested by the reanalysis; quite a large assumption!

behaviour of the LMM, however for an initial exploration and development of methodology here the use of the 20th century reanalysis dataset is appropriate.

Three regions were chosen for their unimodal rainfall climate (i.e. they have a clearly defined single rainy season). For each, ten gridpoints showing a similar seasonal cycle were selected, in order to inflate the LMM input sample size. This gave 140 years x 10 gridpoints = 1400 separate years of temperature/rainfall time series, which were then individually used to drive the LMM. A map of the latitude stripes is shown in figure 8.1, and the climatology of each is shown in figure 8.2. Rainfall seasons for each region are defined as A, B and C.

Looking at the different climatologies, regions A and B show a Sahelian climate, with reduced rainfall for the higher latitude region. For region C the climate is unimodal, describing the southward progression and return of the inter tropical convergence zone. Figure 8.2 shows that the grid points within each region come from a sufficiently unimodal distribution.

The time series for each region was then used to drive the LMM. The seasonal cycle of malaria climate for the three regions is shown in figure 8.3.

Based on figures 8.2 and 8.3 the climate and malaria seasons were defined as JJAS and SOND for the Sahelian regions (A & B), and DJFM and FMAM for region C[2]. Using these definitions, parameters were chosen to describe the season. Climate parameters were calculated based on common statistical properties, and malaria parameters were simply defined as the average values of the standard output of the LMM over the malaria season. A list of parameters and description is contained in table 8.1. These parameters were then compared to see how these average properties of the rainy season co-vary with each other, and with average properties of the malaria season.

Impact surfaces were then created, by looking at the variation in LMM output when driven by input with similar average properties. Average temperature and precipitation were chosen as 'predictors' and incidence as a 'predictand'. After separating the input into an arbitrary number of bins (20 were chosen for each dimension), the mean and standard deviation of incidence was calculated for each bin. The mean gives the expected average malaria incidence following a season with a certain average climate, whilst the standard deviation indicates the variability of LMM behaviour and gives a measure of the relative uncertainty associated with certain climate average states over others. Finally the number of points within each bin is used to mask the mean and

[2]Four month rather than three month seasons were chosen to minimise the variability between time series due to effects outside of the season definitions.That is, to more fully capture any effect of sub-seasonal climate variability on malaria output: the smaller the window, the smaller the fraction of variability explained.

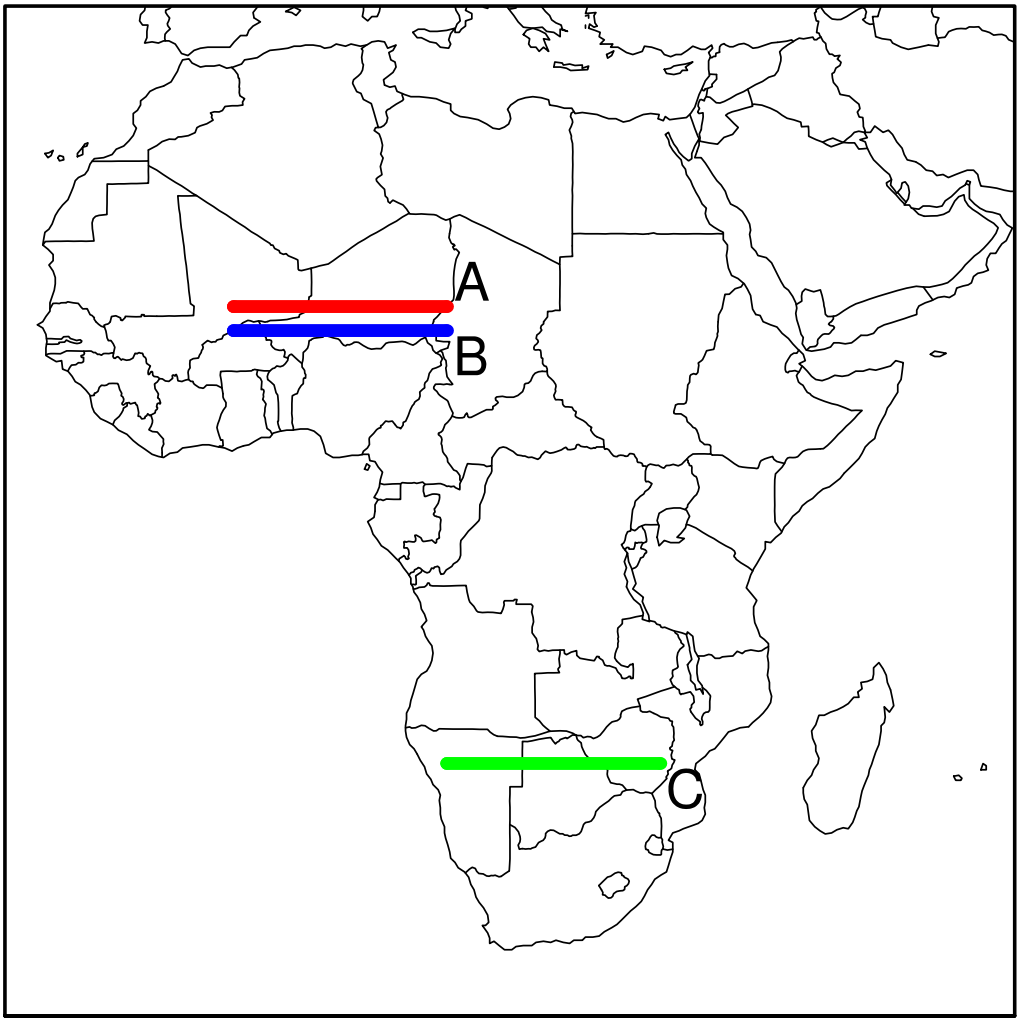


Figure 8.1: Latitude stripes defined as region A, B & C.

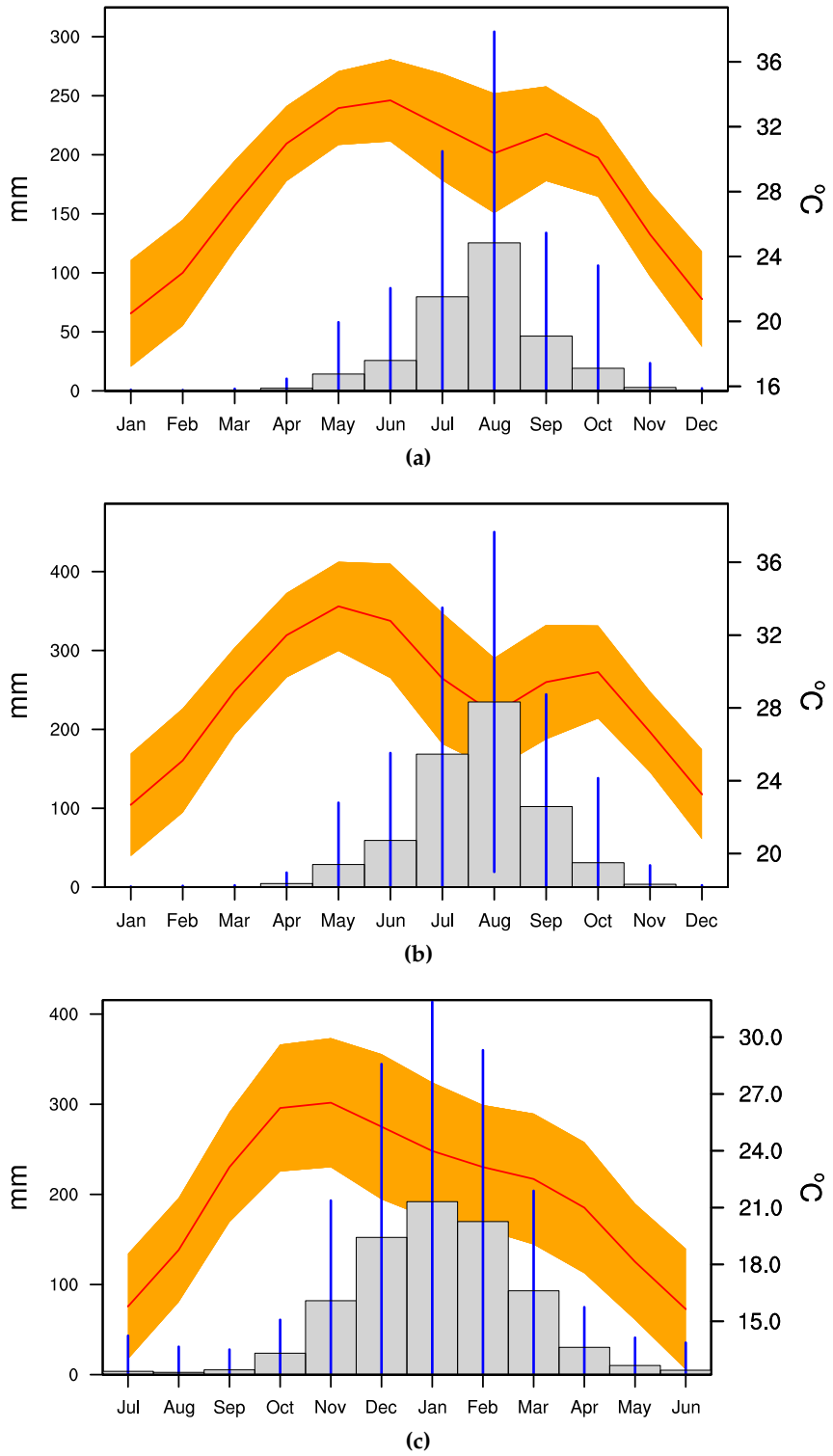


Figure 8.2: Seasonal cycle of temperature and precipitation from the 20th Century reanalysis dataset for the subset of grid points shown in figure 8.1. Warm colours indicate temperature seasonal cycle with 5-95 percentiles, and grey bars indicate rainfall cycle, with 5-95 percentiles.

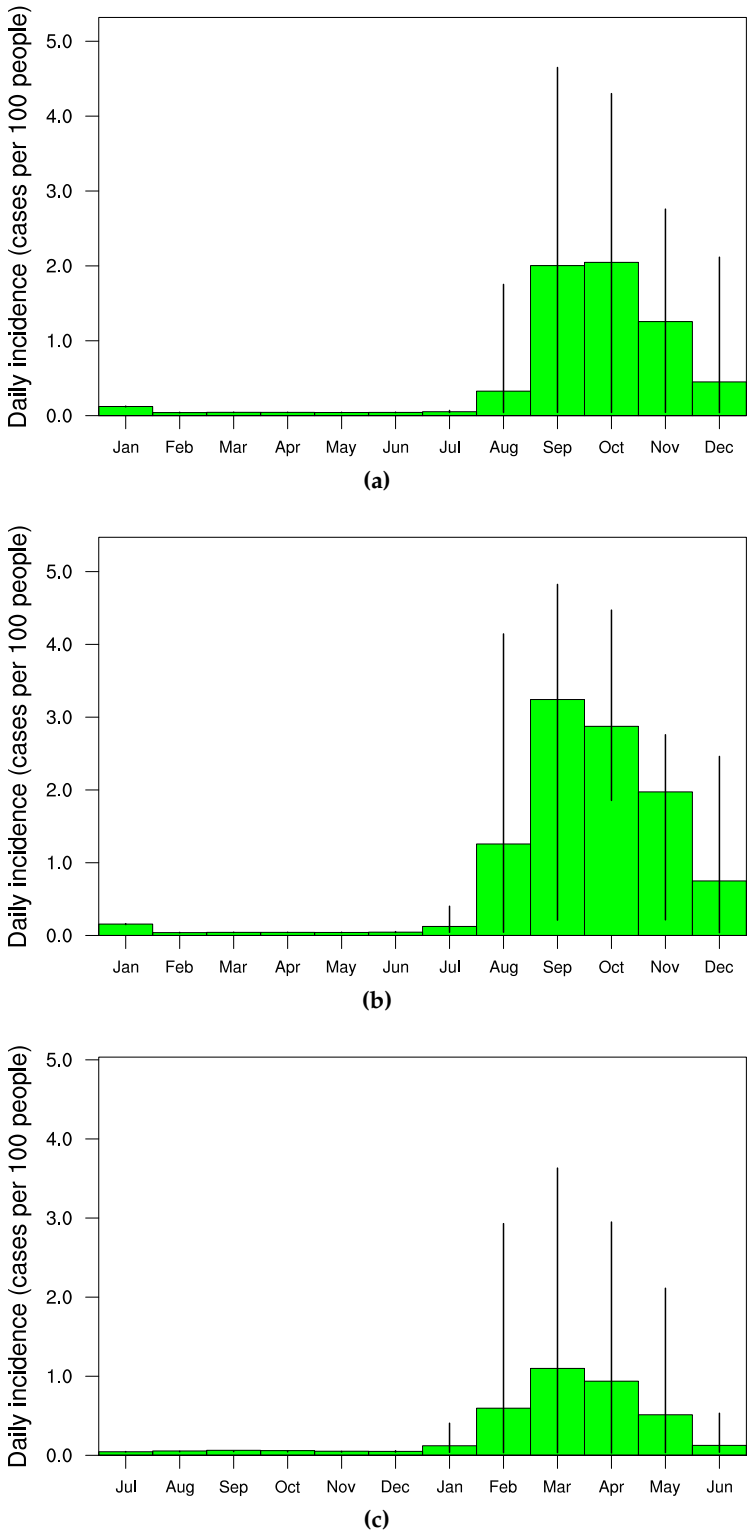


Figure 8.3: Seasonal cycle of LMM incidence when driven by the 20th Century reanalysis dataset for the subset of grid points shown in figure 8.1, error bars show 5-95 percentiles.

Short name	Type	Description
Avg T	Climate	Average temperature
Total rain	Climate	Total rainfall
SD of T	Climate	Standard deviation of temperature
SD of rain	Climate	Standard deviation of rainfall
DD > 18	Climate	Degree days over 18°C
Heavy rain days	Climate	Number of days of rain > 10mm
Rain days	Climate	Number of days of rain > 1mm
Min T	Climate	Minimum temperature
Max T	Climate	Maximum temperature
# breaks	Climate	Number of breaks in rain; defined as at least three days with rain under 1mm
Incidence	Malaria	Average malaria incidence (cases per 100 people)
Prevalence	Malaria	Proportion of the human population infectious
# mature mosquitoes	Malaria	Number of mature mosquitoes
# inf mosquitoes	Malaria	Number of infectious mosquitoes
# act mosquitoes	Malaria	Number of active mosquitoes
G Days	Malaria	Average length of the gonotrophic cycle in days
S Days	Malaria	Average length of the sporogonic cycle in days

Table 8.1: A list and description of the climate and malaria season parameters used. Climate parameters were chosen from common statistical measures, whilst malaria parameters are seasonal averages of the standard output of the LMM.

standard deviation points. Where the number of points is less than three the corresponding bin is not included on the mean and standard deviation surfaces (due to the unreliability of estimating the variance and mean in LMM behaviour from only a few points).

These impact surfaces can then be used to explore malaria model uncertainty. There is uncertainty in parameters and structure, and using different realisations of the model to create impact surfaces gives an idea of how sensitive the surfaces are to changes in the model. A large source of uncertainty in the behaviour of the LMM is the choice of survival scheme (Jones, 2007). The survival scheme is the mathematical formulation within the model which describes the dependence of the mosquito mortality rate on temperature. Four schemes are currently built into the LMM: Martens (Martens et al., 1995), Lindsay/Birley (Lindsay and Birley, 1996), Bayoh (Bayoh, 2001) and the Craig version of Martens (Craig et al., 1999). Details of these schemes can be found in the relevant papers and will not be discussed further here. Hereafter they will be referred to as schemes one, two, three, and four[3].

[3] All other results using the LMM in this thesis LMM have used the Martens survival scheme, which is generally used as the default

Each survival scheme produces a different impact surface for each region. How this uncertainty information can be combined is then explored - description is deferred to the corresponding results section.

8.3 Results

8.3.1 Relationships between low temporal resolution climate and malaria parameters

Firstly turning to the scatter plots showing the relationship between the parameters defined in table 8.1. For brevity only results for region A are shown. Figure 8.4 shows how the climate parameters covary with each other for JJAS averages.

There are some obvious relationships visible for the parameters related to temperature. For instance, there is a positive correlation between average and maximum temperature, and between average and minimum temperature. There is also a strong correlation between degree days over 18°C and minimum, maximum and average temperature. Interestingly however maximum and minimum temperature are not strongly related.

There is no clear relationship between standard deviation of temperature and average temperature, though there is a strong negative correlation between minimum temperature and standard deviation of temperature. There is also a weak positive correlation between maximum temperature and its standard deviation.

For the rainfall variables there are also some clear relationships. For instance there is a strong correlations between total rain and heavy rain days/rain days. There is also a positive correlation between total rainfall and the standard deviation of rain - unsurprising since total rainfall is bounded by zero, so situations when the total rainfall is higher are likely to have a higher range and on average a higher standard deviation. For breaks in the cycle there are some areas of the parameter space not covered, for instance the high rain day/high break space, though this is a trivial result; when more days in a season have rain, it follows that there will be fewer stretches with no rain. N.B. The plots for number of breaks appear different to the others since the variable is an integer and has a relatively small range.

Looking at the relationships between temperature and rainfall variables, it can be seen that generally there are negative correlations. For example, between average temperature and each of total rain/rain days/heavy rain days/standard deviation of rain the correlation is negative. This is likely due to the cooling effect of rainfall on the

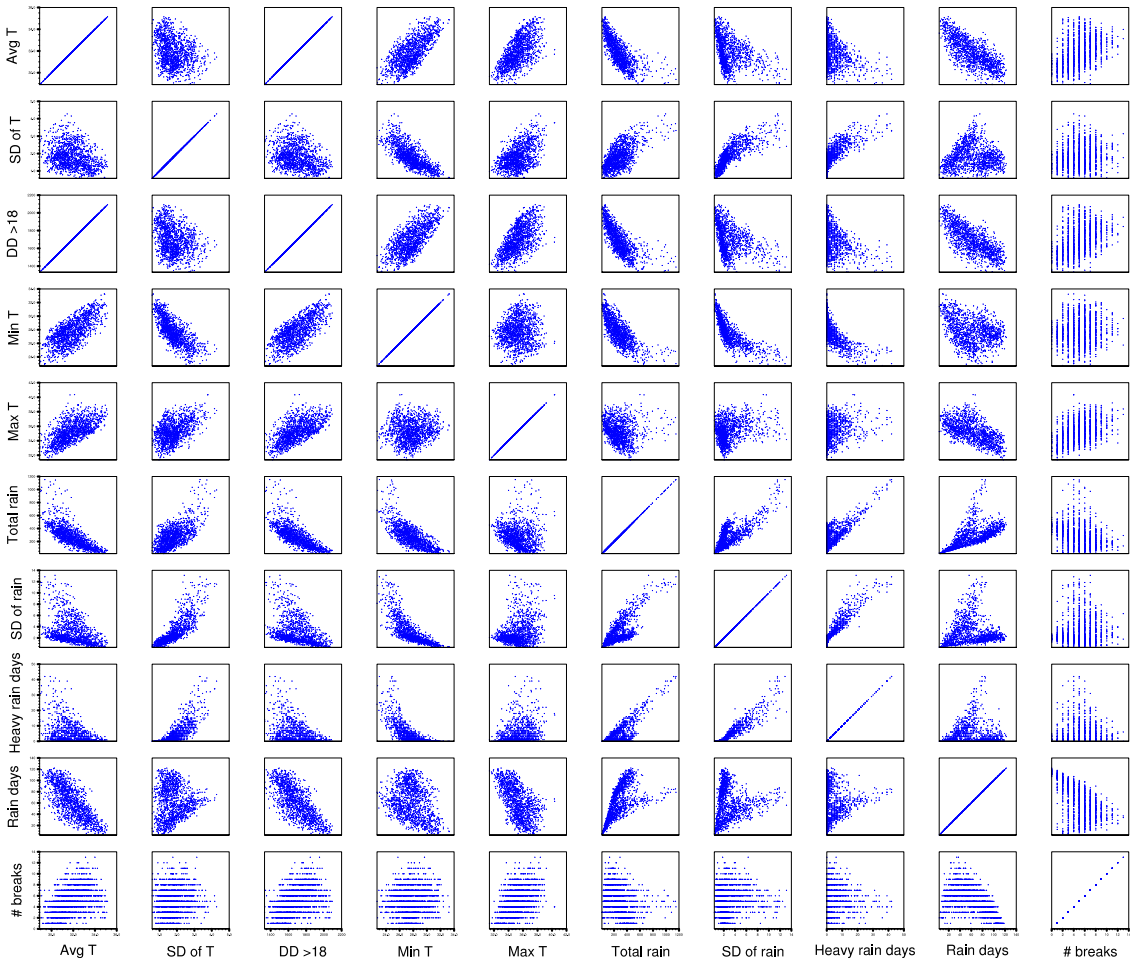


Figure 8.4: Scatter plots for climate parameters vs climate parameters, JJAS, region A.

surface, thus when a season has high (low) rainfall the average temperature will be accordingly low (high). There is also a strong positive correlation between the standard deviation of temperature and standard deviation of rainfall, which presumably is driven by rainfall; when a season has high rainfall variability, the cooling effect from the rain occurs on some days and not others, giving a corresponding high variability in temperature.

A final point to note for this plot is that some plots have a double shape (for instance between total rain and rain days), as if the points come from at least one separate distribution. This suggests that the assumption that the points in the latitude strip come from an identical climate is not entirely valid.

Scatter plots of SOND malaria parameters for region A are shown in figure 8.5. Here there are clear positive correlations: between incidence and prevalence, and between numbers of infective, mature, immature and active mosquitoes. The relationship

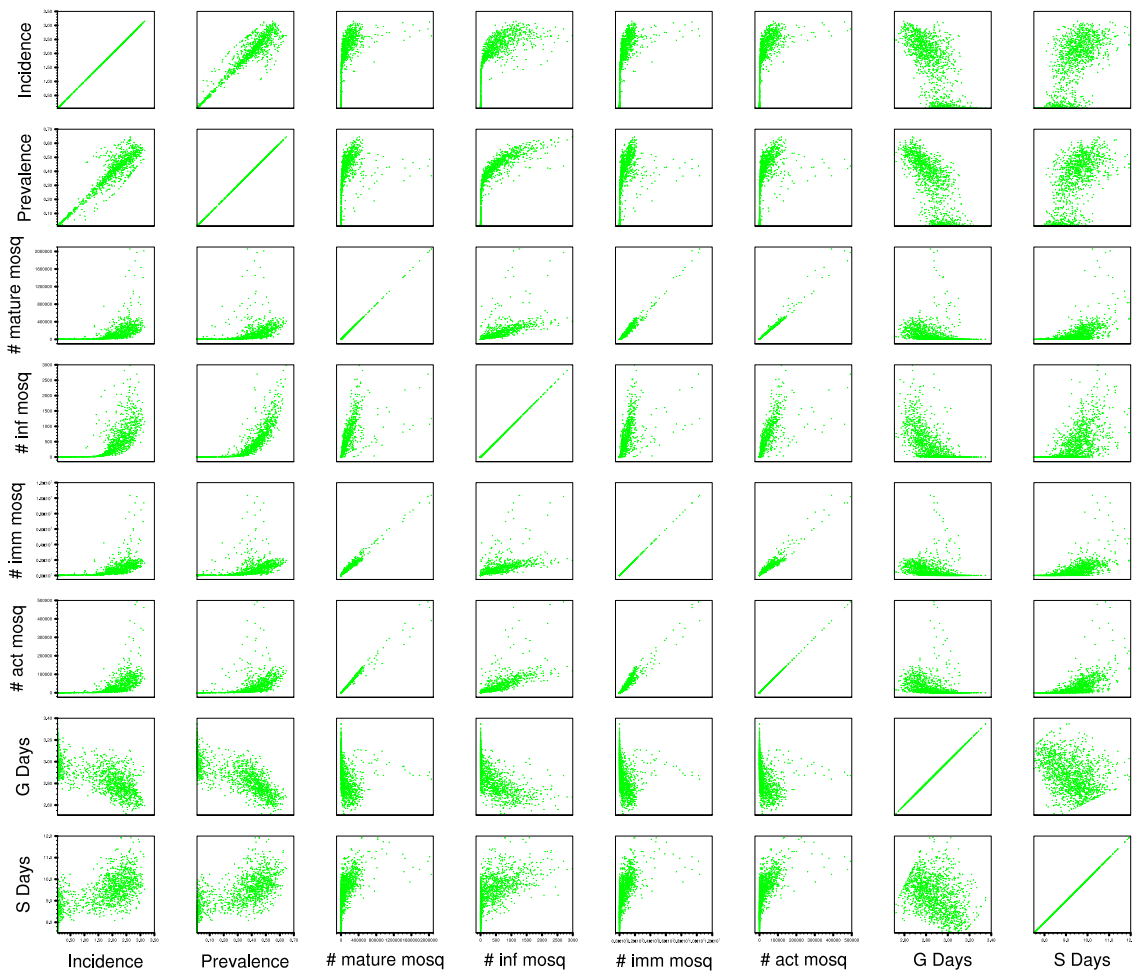


Figure 8.5: Scatter plots for malaria parameters vs malaria parameters, JJAS/SOND, region A.

between the numbers of different kinds of mosquitoes with one another is also generally positive. The average relationship between gonotrophic days and incidence is slightly negative, whilst that between incidence and sporogonic days is slightly positive[4].

Finally, relationships between climate and disease variables are shown in figure 8.6. The clearest positive correlations for incidence are between it and rain days and also total rain. There is a slight negative relationship between incidence and average temperature, with all incidence damped to zero above 34°C. There is also a damping of incidence to zero when the number of rain days is below 20.

[4]There is also a clustering of points around zero incidence/prevalence (e.g. gonotrophic days vs. incidence). This clustering is not visible on any of the mosquito number plots; here the number of mosquitoes is zero. This is then the reason why incidence and prevalence is zero. However the reason for the absence of mosquitoes is not clear; it could be an aspect of the climate which causes the LMM to fail, or it could be a technical bug in the model.

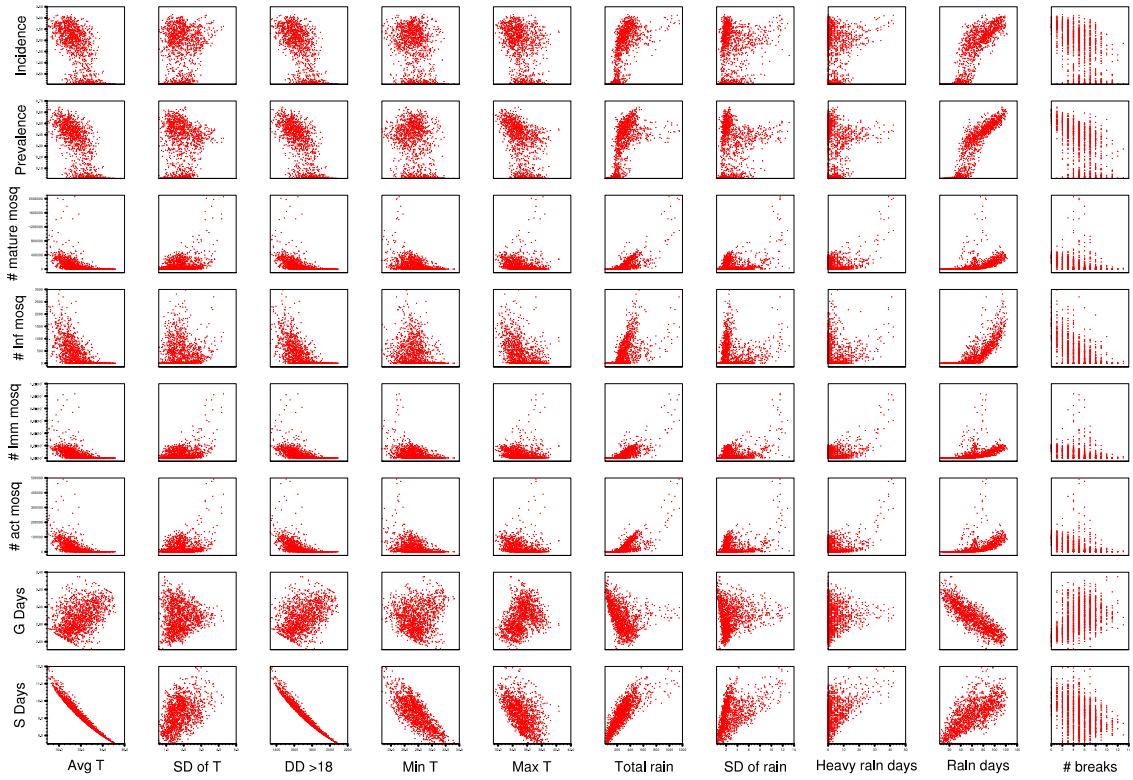


Figure 8.6: Scatter plots for climate parameters vs malaria parameters, JJAS/SOND, region A.

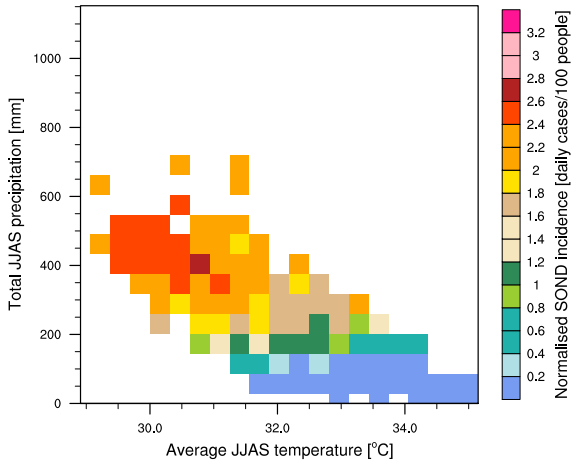
Incidence seems to be the most relevant aspect of the LMM output related to seasonal malaria burden. Furthermore, average temperature and precipitation are generally available as climate forecasts. How these variables relate to incidence as they covary is the subject of the following section.

8.3.2 Impact surfaces

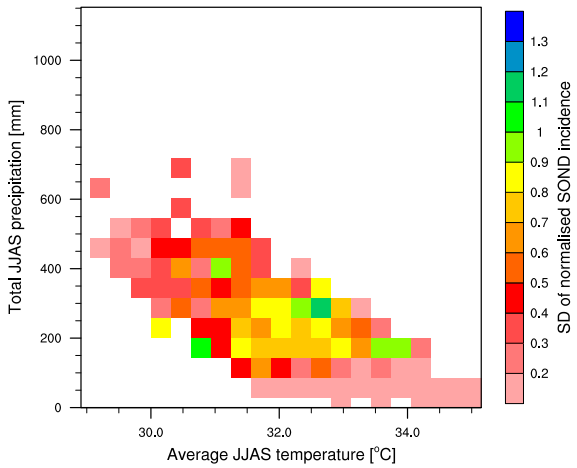
Shown in figure 8.7 are impact surfaces for region A; showing the response of the LMM to different average climate states. There is a clear distinction between high and low incidence (figure 8.7a); when the rainfall is below 100mm total over the season, incidence is low. For lower temperatures and high rainfall in this region the incidence is highest. The standard deviation of incidence is greatest for climate states in the centre of the climate space, and lowest for those around the edge (figure 8.7b). This is not an artefact due to a low mean subset having a corresponding low standard deviation, since the regions with the high mean also have some of the lowest standard deviation. For this region then it suggests that a climate forecast for either end of the distribution can be confidently said to normally relate to a high or low malaria season, whilst states in the centre of the space have a greater uncertainty associated with them.

For region B (the lower latitude Sahalian region, figure 8.8), the mean incidence is much higher than for region A, though the contouring of the incidence surface is similar; high malaria years for low temperature/high rainfall years, whilst low rainfall seasons precede have a relatively low malaria season. The plot of standard deviation however is different; the uncertainty is much lower everywhere, except for a small subset of the space, for low rainfall events. This suggests a lower interannual variation in incidence in this region.

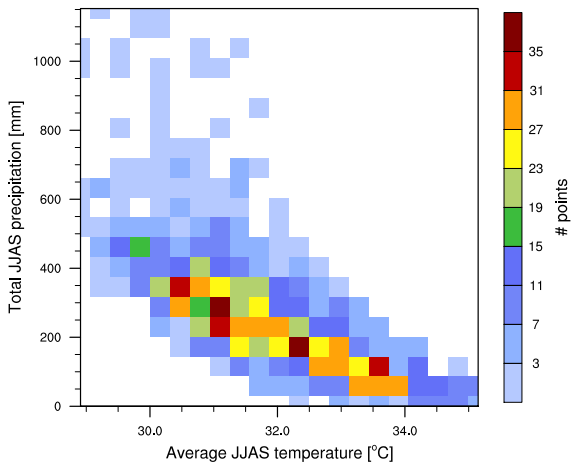
Finally for region C (the Southern African region, figure 8.9), the shape is different, with the lowest temperatures and rainfall relating to low mean incidence and uncertainty in the forecast (N.B. the domain and range of the temperature and rainfall axes are different for each region). Incidence is highest for climate states where precipitation is high and temperature is also high. This is also a region of climate space where the uncertainty is highest; though the high variation is not simply due to the high mean as there are points with a high mean incidence yet low standard deviation.



(a)

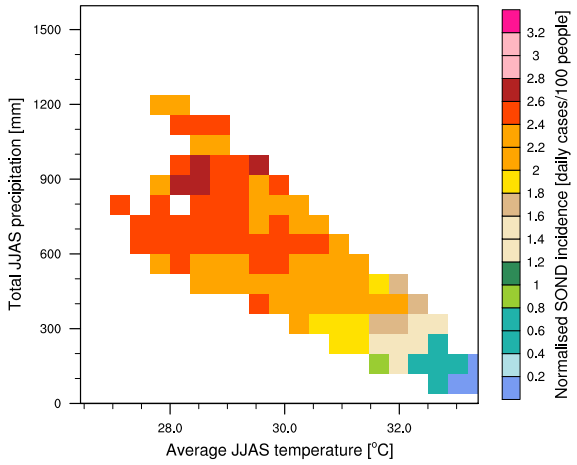


(b)

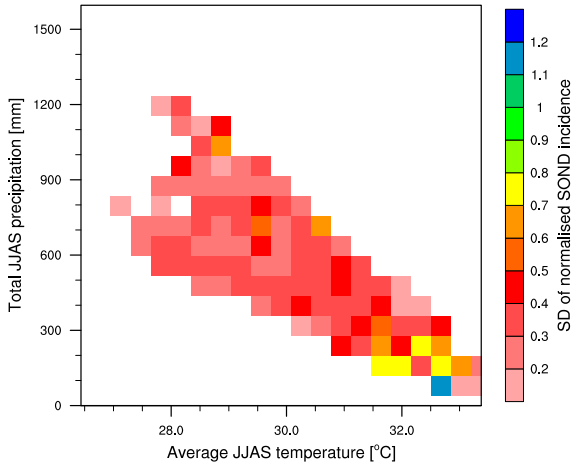


(c)

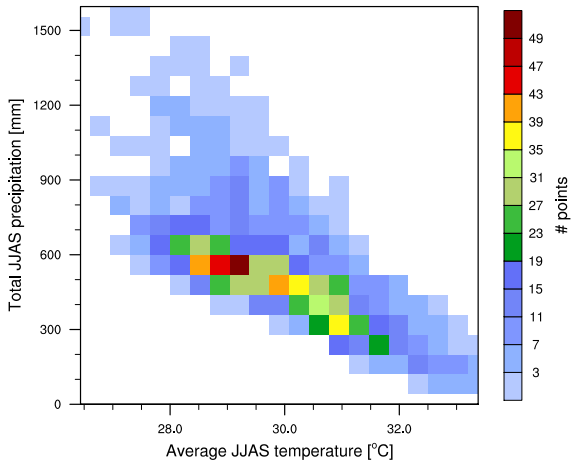
Figure 8.7: Impact surfaces for region A. Showing mean (a) and standard deviation (b) of SOND incidence along with the number of points in each climate bin (c).



(a)

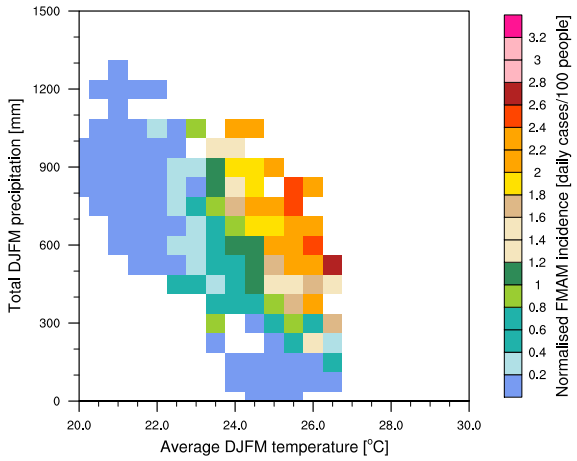


(b)

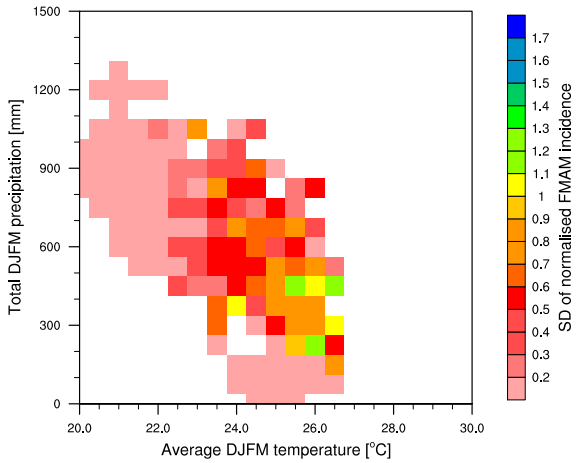


(c)

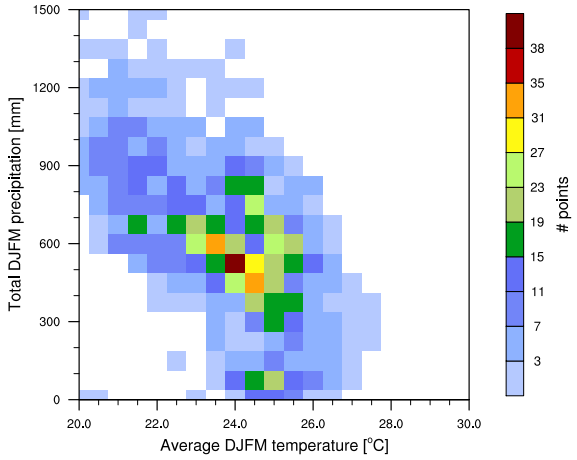
Figure 8.8: Impact surfaces for region B. Showing mean (a) and standard deviation (b) of SOND incidence along with the number of points in each climate bin (c)



(a)



(b)



(c)

Figure 8.9: Impact surfaces for region C. Showing mean (a) and standard deviation (b) of FMAM incidence along with the number of points in each climate bin (c)

8.3.3 Exploring the uncertainty

Now we turn to the effect on LMM when different survival schemes are used. Results follow for region A, with those for regions B and C left to appendix E. Climatologies are shown in figure 8.10. It can be seen that generally the shape is similar between survival schemes, transmission between September and January and low incidence outside of this period. The magnitude of incidence between survival schemes however is quite different, with scheme 1 showing the lowest incidence and the incidence for scheme 4 peaking early and remaining higher for longer.

Mean incidence and its standard deviation are plotted in climate space for the survival schemes in figure 8.11. Generally the magnitude is different between the plots of mean, whilst the relative contouring of the space is similar; the low temperature/high rainfall space has highest incidence, whilst the high temperature/low rainfall space has low incidence. There is a significant difference in the uncertainty information[5]; for survival scheme 1 the most uncertain area is in the top left of the space (figure 8.11b), whilst it is in the centre of the space for 2 and 3 (figures 8.11d and 8.11f), and for scheme 4 it is closer to the bottom right of the space (figure 8.11h). This shows that the use of one scheme can provide quite different information to another; if only one was used then very different confidence in the relationship between average climate and malaria incidence might be assumed. If all schemes (and by extension, parameters and also malaria models) are valid, then all should be included in the analysis for a full quantification of uncertainty.

How then to combine this information? The distribution of the LMM output for each survival scheme is clearly different, that is, over the whole climate space the mean and standard deviations are different, preventing a simple combination of the information. Potentially incidence for each scheme could be normalised and combined, however the differences in standard deviation between survival schemes would remain.

To combine the information, tercile categories were employed. After calculating the 33% and 67% thresholds, each of the individual 1400 points was classified as belonging upper, middle or lower tercile category for each survival scheme separately. The data was then combined and simplified following a set of rules. For each of the climate bins, the subset of points within was summarized as follows. Bins with more than 50% of members in the upper tercile category are described as 'upper tercile'. Of these, those with both more than 75% were in the upper tercile category and more than ten members in the bin are put in the 'upper tercile: high confidence' category. Others are described as 'upper tercile: low confidence. This reflects statistical significance in a general way - where the

[5]As discussed previously, standard deviation is used here as a proxy for the uncertainty in LMM behaviour associated with certainty climate states over others.

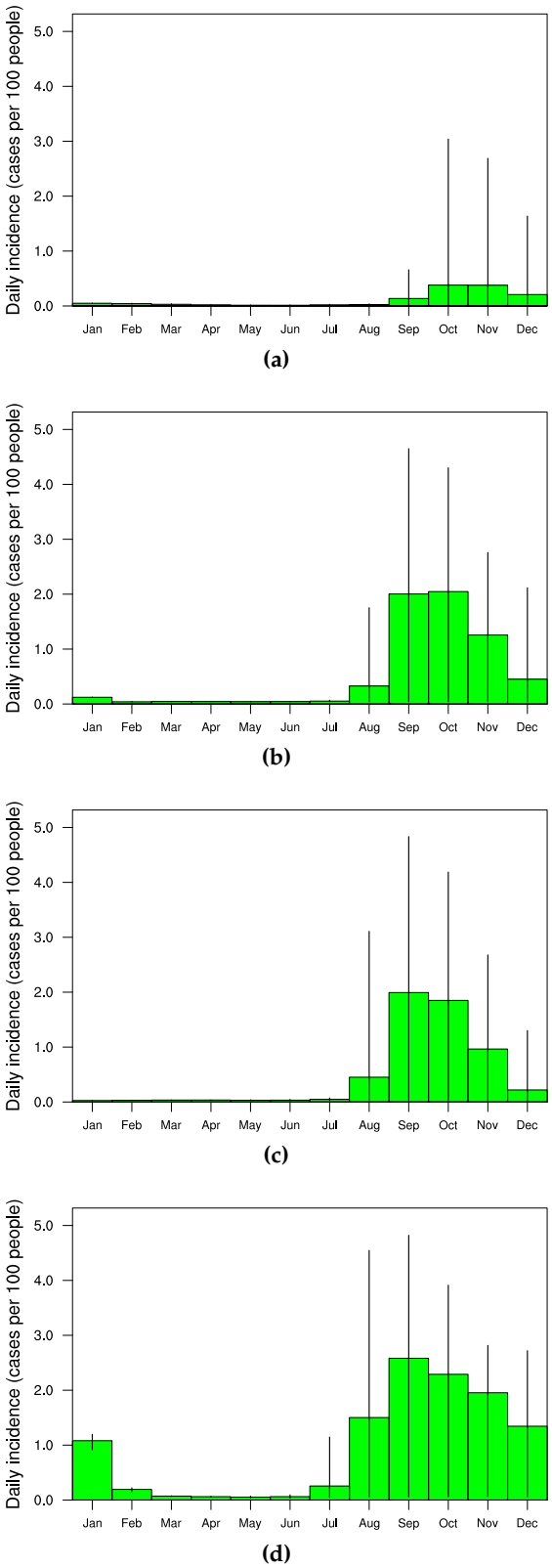


Figure 8.10: Seasonal cycle of LMM incidence when driven by the 20th Century reanalysis dataset for region A, using survival schemes one to four (a-d).

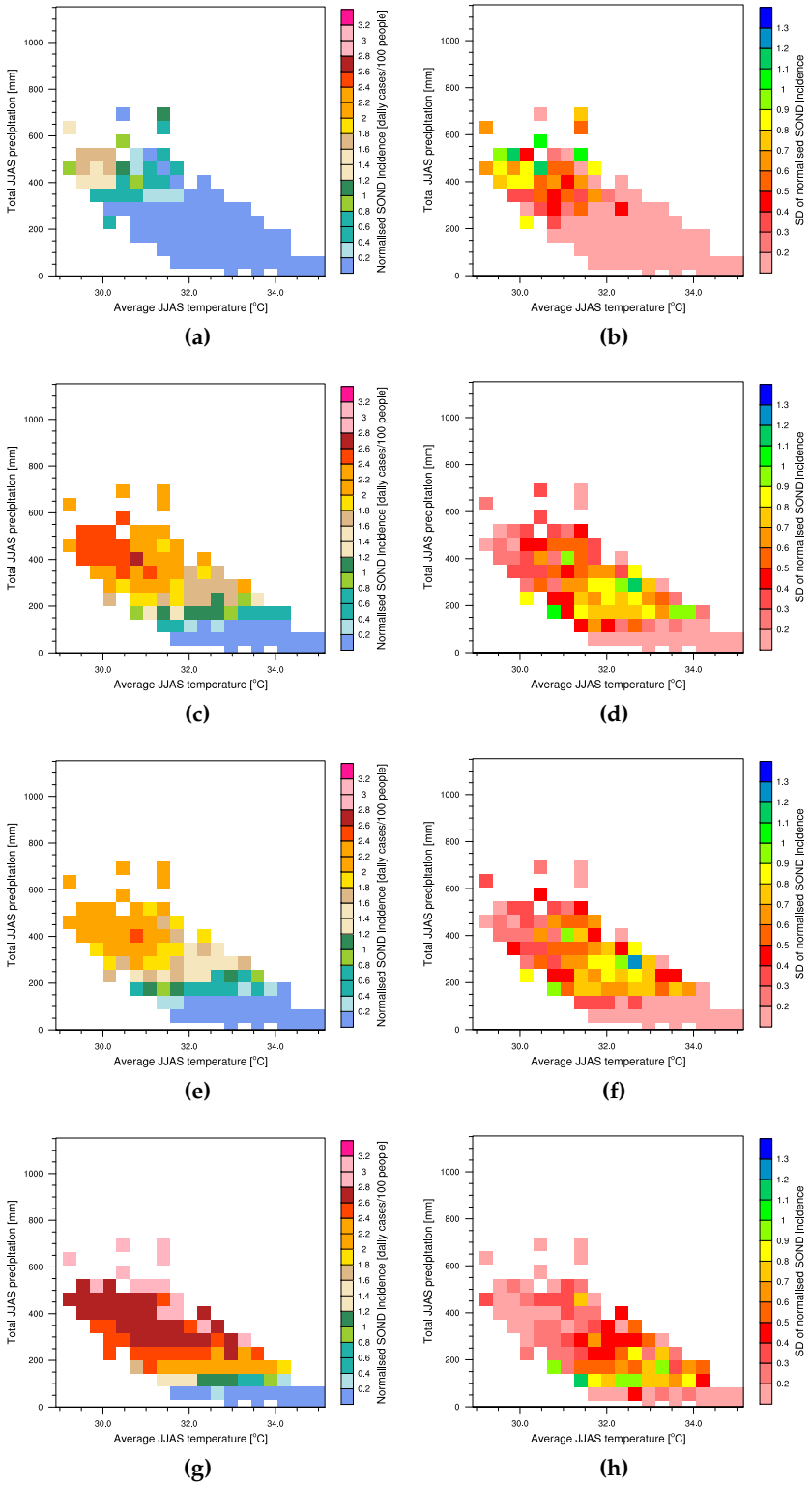


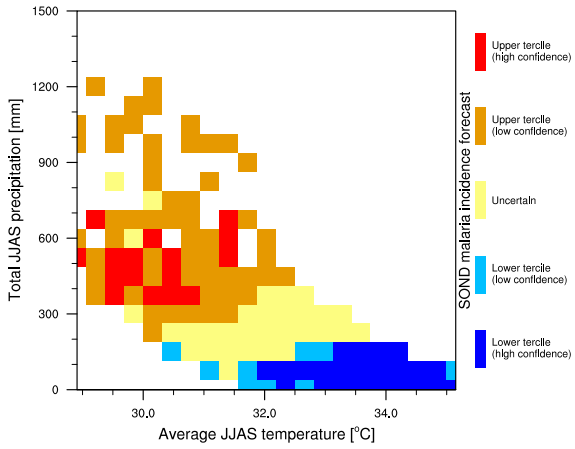
Figure 8.11: Impact surface comparison, mean (left) and standard deviation (right) for region A. Using survival schemes one to four (top to bottom rows).

sample size is low a relationship between average climate and disease cannot be stated with confidence[6].

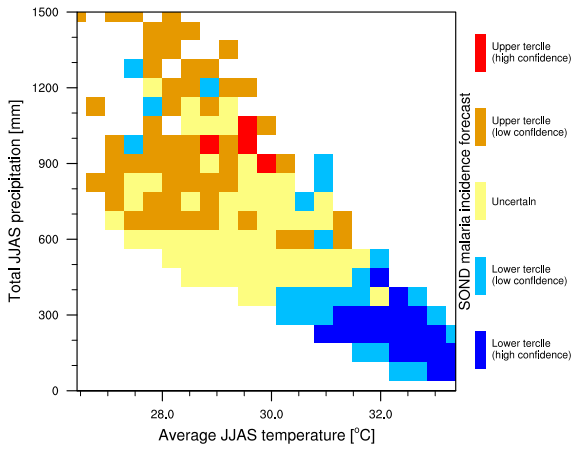
The converse was done for lower tercile events, classifying bins into either 'lower tercile: high confidence' and 'lower tercile: low confidence' based on the fraction of members in the lower tercile category. All climate bins outside of these four descriptions are then defined as 'uncertain'. This includes all which have a majority of middle tercile members. The rationale behind this is that a 'high confidence middle tercile' forecast essentially means that the average of climatology is expected, and advice given based on this would be along the lines of 'do what you would normally do'. This is not particularly useful; in this case perhaps it is best that no strong advice is given. An early warning forecast system of this nature should perhaps only be used for anomalous (i.e. upper and lower tercile) events.

These rule-combined impact surfaces for each region are shown then in figure 8.12. The shape for each region generally follows the mean incidence pattern, though the graphic integrates uncertainty information from each survival scheme. It is easy to see for each region the climate states when action may be taken and others where the outlook is not as certain.

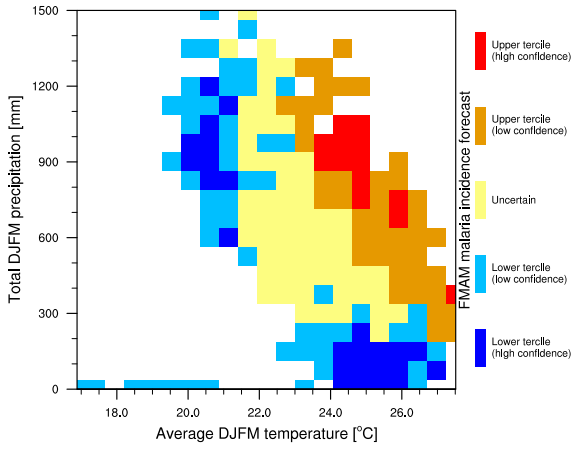
[6]Thresholds of 50%, 75% and ten members are chosen based on intuition, and certainly the case for different choices could be argued. Different end-users of risk information may prefer different thresholds; for instance higher thresholds would be more appropriate for a more risk-averse user. However for an initial exploration and design of graphic the choices made here are appropriate.



(a)



(b)



(c)

Figure 8.12: Impact surfaces, combining all survival schemes, in average temperature total precipitation space, for regions A, B and C.

8.4 Discussion

The impact surfaces created here at the culmination of this investigation are an effective way of communicating the malaria risk and uncertainty due to climate drivers. They can be potentially also be combined with a climate forecast, this is shown in figure 8.13. Individual climate forecast ensemble members can be mapped to the surface, with their positions updating over time. Due to the lag between climate and malaria, as the rainfall season progresses and concludes, the ensemble of climate simulations will evolve into a single point.

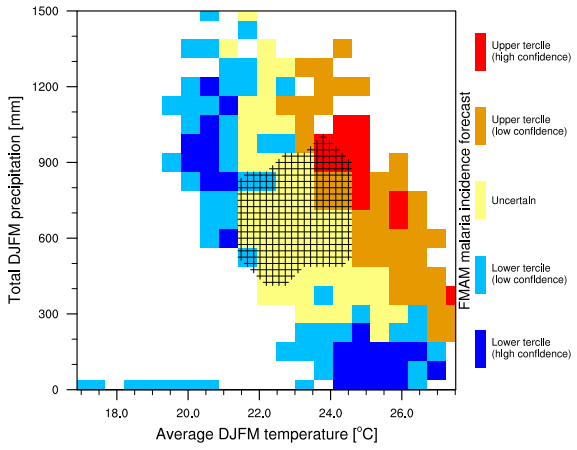
Combining a climate forecast with an impact surface shows the confidence in disease outcomes intuitively and how the forecast will update over time. The relatively simple end-product rests upon complex modelling and quantification of uncertainties, allowing the effective communication of information related to forecast confidence.

The method does not require any creation of a smoothed probability distribution function before use; climate input can essentially be used raw. It would likely need bias correction - a seasonal climate model will not have an identical climate as the reanalysis used here. The bias correction however would only be necessary on monthly and longer averages, relatively easy when compared to bias-correcting daily data, normally required for LMM. Furthermore, the need for bias correction could be circumvented entirely by choosing the bins when creating the impact surface based on percentiles. All that then would be needed is climate ensemble forecast populations within tercile bins based on the forecast model's own climate distribution.

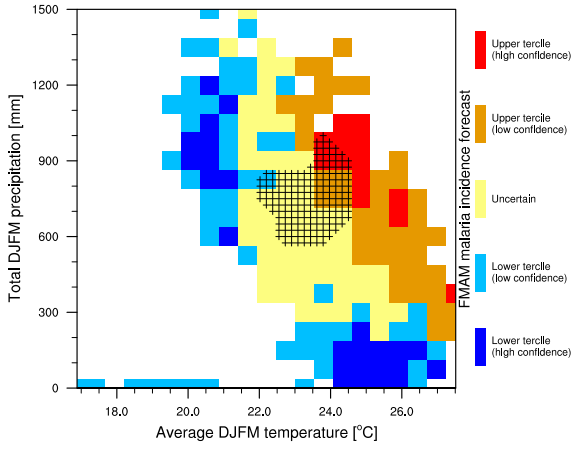
Generally the impact surface patterns are coherent though occasionally there are points which appear in unexpected places (i.e. single points for low tercile close to high tercile areas). The reason for this could be due to the sample size in an individual bin, and a larger sample of points would test to see if these are robust (i.e. by using a weather generator, see below).

The climatology is different when using different survival schemes. Whilst generally the shape is similar, with the malaria peaking around the same few months regardless of the choice of survival scheme, the magnitude of the incidence is much higher for some schemes than others. Furthermore some survival schemes have a higher incidence in earlier months outside of the targets used here. This suggests that perhaps a longer target may be more appropriate.

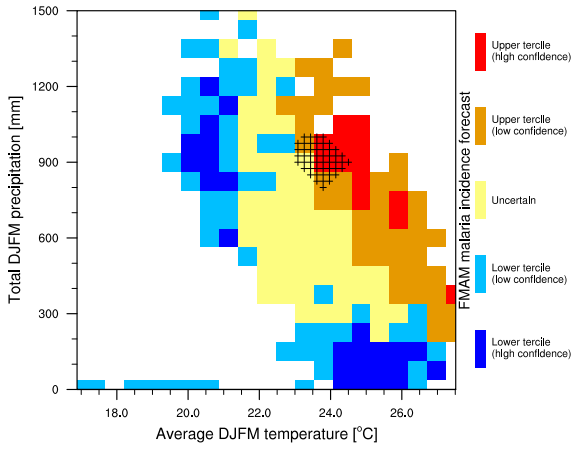
All survival schemes have been treated as equal here, though with expert opinion and consideration perhaps some could be weighted or eliminated as unrealistic for



(a)



(b)



(c)

Figure 8.13: Impact surfaces with theoretical ensemble climate forecasts superimposed as hatching, moving from a wide ensemble spread (a), through a medium spread (b) to a narrow range of possible climate futures (c), facilitating an intuitive understanding of quantified uncertainty in malaria risk.

particular regions. This work is however an initial exploration and proof of concept; further investigation would no doubt improve accuracy of the plots.

There is also an assumption about the inputs: that the daily time series of the reanalysis dataset employed is representative of reality. No analysis of the realism of the sub-seasonal variability in the reanalysis was carried out; it was used in an *ad hoc* manner. This assumption could be investigated by comparing the reanalysis to station data in relevant regions. The investigation could also be repeated by using synthetic data generated, by for instance a weather generator (e.g. Richardson et al., 1998).

Furthermore, analysis of the covariance of the climate parameters has shown that the input data potentially comes from more than one climate distribution (figure 8.4, e.g. between total rain and rain days). As such the assumption that the grid points used in each sample came from the same distribution is questionable. Synthetic data may therefore be more appropriate for an unbiased test of LMM behaviour.

Uncertainties related to the model structure of the LMM have only been explored here in a limited way; a more detailed investigation is warranted. Instead of using different survival schemes, model parameters could be varied in order to explore the uncertainty in each (similar to climate model perturbed parameter experiments). Co-varying these parameters would also further explore the combined errors, working toward a goal of fully quantifying the uncertainty within the LMM. This method would easily allow the incorporation of further malaria models such as LMM_{2010} (Ermert et al., 2011a) and VECTRI (Tompkins and Ermert, 2013), in a multi-disease model approach toward uncertainty quantification.

The question remains; from where does LMM output variability arise? The source could be in characteristics of the input outside of the rainfall season defined here. It is also likely to arise at least partly from characteristics of the input time series within in the rainy season, i.e. the sub-seasonal variability. The distribution of rain throughout the season is likely to be important, as is the phase angle between peaks in temperature and rainfall. The effects of these could be explored by by comparing the output from an 'ideal' rainfall synthetic time series with multiple series where the peak is shifted sequentially forward and backward in time and the sub-seasonal variability modified.

The effect of noisy input on the output is another important question: how much noise can be added to temperature and rainfall time series before incidence is changed significantly? This is an important question if the LMM is to be used with seasonal climate models, where there is likely to be a significant amount of noise in forecasts. Extending the methodology in this chapter to use a weather generator would create a deeper understanding of the behaviour of the LMM. Parallel to this, developing the

method to include variation of multiple parameters and alternative malaria models would allow a fuller quantification of disease model uncertainties, ultimately creating a useful decision-making tool.

CHAPTER 9

Discussion and conclusions

A discussion concludes this thesis, split into four sections. Firstly the work carried out and the main conclusions are summarised for each chapter separately. The main results are then discussed in relation to decision making. Following this limitations of the work and avenues for further work are described and finally the chapter concludes with a short discussion of uncertainty in prediction of climate-driven disease risk relating to climate change.

9.1 Summary of work

Part 1 focused on validation of climate models. Results presented in chapter 4 looked at decadal prediction, analysing the hindcasts produced as part of the ENSEMBLES project. Some skill in the prediction of global average temperature trends over the forthcoming decade was found, with no skill for precipitation. Analysis focusing on local subregions found limited skill in predicting temperature trends for these regions, with again no skill for precipitation. The main conclusion of this chapter is that decadal models are currently not skillful enough to make useful predictions of upcoming disease risk.

Chapter 5 looked at the evolution of seasonal forecasting skill, and compared the hindcasts from two research projects DEMETER and ENSEMBLES (from 2004 and 2008 respectively), with the most recent version of the ECMWF seasonal forecast model, System 4. Models were validated over Africa and the Indian subcontinent, and System 4 was found to generally provide the most skilful forecasts. Temperature and precipitation biases are generally lower in System 4 than in ENSEMBLES and DEMETER, whilst the areas of significant ensemble mean correlation and relative operating characteristic area under curve (ROC AUC) are largest for the more recent model. Potential economic value of forecasts was also described and is generally highest for System 4, particularly over West Africa. Results summarising the

improvement in brier skill score and economic value can be found in table 5.5; the main conclusion of this chapter is that there has been an improvement in seasonal climate prediction skill over the past decade, and System 4 is an optimal choice for driving a disease model.

Subsequently, chapter 6 took an in-depth look at System 4, comparing the variation in skill between forecast start dates. Potential economic value was demonstrated at multiple lead times (results are summarised in table 6.2). Most skill was found for west African regions and Botswana, whilst skill of System 4 over the Indian subcontinent is limited. Furthermore, skill and its evolution as a target is approached generally increases, but it can also decrease; therefore one cannot assume that the closest forecast to a target will be the best one.

Part 2 then turned to the question of the potential for climate-driven disease prediction. Malaria prediction was studied using the Liverpool Malaria Model (LMM). Chapter 7 looked at the skill of LMM output when driven with System 4 hindcasts, firstly validating the model over Botswana where malaria data is available for validation (in the form of the Botswana Malaria Index (Thomson et al., 2005). Here skill was found, with positive economic value above 95% significance. Value and ROC AUC was highest for November System 4 start dates, showing an improvement over previous work driving the LMM with the DEMETER seasonal hindcasts issued from the same time (Jones and Morse, 2010).

This tier-3 validation was followed by tier-2 validation of three African regions where System 4 demonstrated skill in chapter 6: the Sahel, the Gulf of Guinea and Malawi. For these regions malaria data is not available for tier-3 validation, instead validation was carried out using the LMM driven by the ERA-Interim reanalysis as a reference. Skill was found at the simulated epidemic fringe of the Sahel, and in north west Malawi. The Gulf of Guinea showed no skill at tier-2 validation. This is consistent with previous work with the LMM showing that the model performs poorly in regions (such as the Gulf of Guinea) where malaria transmission is year-round; poor performance here is likely to be due to the lack of immunity in the malaria model. However since validation is at tier-2 level then this places a major caveat on this conclusion; any problems the LMM has with endemic areas will affect the simulated malaria in both the forecast-driven and reanalysis-driven runs. Finally, tier-2 validation was then followed by the description of a novel method for interpreting tier-3 and tier-2 hindcast validation as a quantification of uncertainty in prediction of climate driven disease risk.

Part 2 and thesis results concluded with chapter 8. This chapter considered the uncertainty in how low temporal resolution climate information can be related to malaria risk. By using the 20th century reanalysis dataset, the relationship between

seasonal average climate parameters was explored, and impact surfaces were created. These relate average temperature and precipitation over a season to the average seasonal malaria incidence in a visually appealing way, and are a tool potentially useful to decision-makers. The robustness of these impact surfaces was investigated, in an initial attempt at quantifying malaria model uncertainty, comparing the impact surfaces calculated when different LMM survival schemes were used[1]. A method of combining impact surfaces based on tercile categories was described and implemented. Finally it was demonstrated how this graphic could be integrated with a seasonal ensemble forecast system, allowing an intuitive visual communication of the uncertainty in prediction.

9.2 Relating results to the decision making process

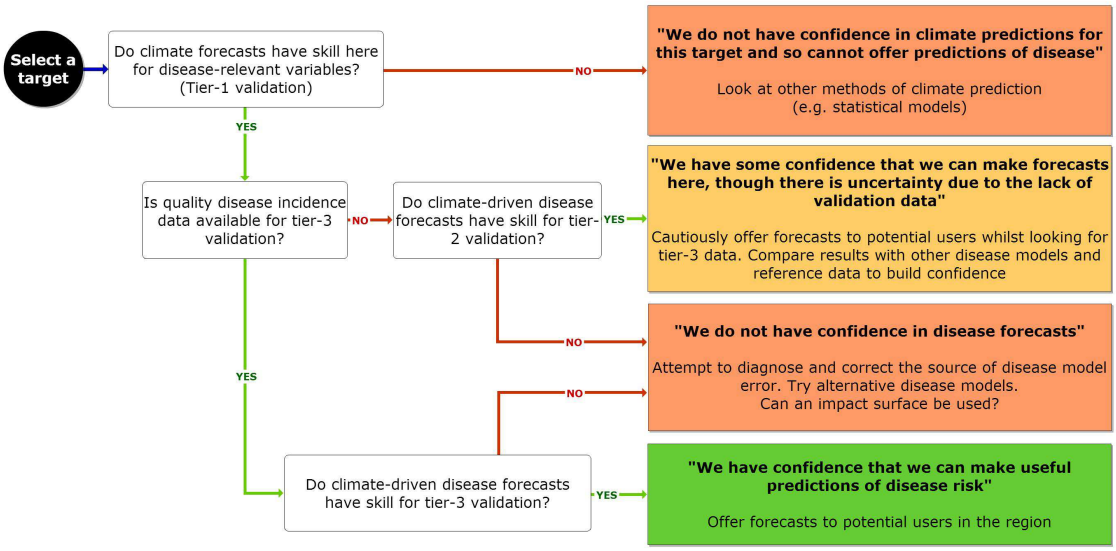


Figure 9.1: A flow chart describing how the validation of climate and disease models can be communicated to decision makers.

In an attempt to communicate how these conclusions can be related to decision making, a flow chart is presented in figure 9.1. Each chapter of research relates to part of this diagram. Firstly a target region is defined, and the first question to ask is if climate models can provide skilful predictions here; part 1 of the thesis determines the route from this box. If the answer is 'no' (as it is for decadal forecasts from ENSEMBLES) this leads to the conclusion that we cannot currently offer forecasts of climate-driven disease risk at this timescale. The direction of research action in this case should be to look at

[1]The choice of survival scheme being one of the more uncertain components of the LMM.

alternative methods of climate prediction, for instance by using statistical models (Dool, 2007), or model output statistics (Glahn and Lowry, 1972).

If climate models have skill at tier-1 validation (as System 4 does), the next question asks whether tier-3 data is available. Epidemiological data of varying quality exists, and so obtaining this data and using it to validate disease models is a priority. If the answer to this question is 'yes', then tier-3 validation can be carried out. In the case of malaria over Botswana tier-3 data is available, this has been used and the LMM driven by System 4 shows skill against this for November start dates. This suggests that these forecasts can begin to be issued on an annual basis to decision makers.

If no tier-3 data is available then validation is only possible at the tier-2 level. Skill at tier-2 level should then be investigated and where it is present (here the epidemic fringe of the Sahel around the border of Mali and Mauritania, and over north west Malawi), forecasts can cautiously offered to decision makers (communicating caveats relating to lack of validation data). Before this, confidence in results should be built by using alternative reanalysis references for validation; results can be also compared with different malaria models. Of course, tier-3 data in these regions would be ideal: areas of skilful tier-2 validation then indicate priorities in the search of tier-3 data.

If there is no tier-2 skill then the conclusion is the same as if there is no tier-3 skill: we do not have confidence in disease forecasts. Research then should focus on diagnosing and correcting the source of disease model error. In the case of the Gulf of Guinea this could mean introducing an immunity component to the model improving simulation in endemic areas. Alternatively another climate-driven malaria model could be employed.

Work contained in the final chapter, relating to impact surfaces, does not fit neatly onto one of the question boxes in figure 9.1. It most closely relates to the top conclusion box: since impact surfaces can be linked with statistical models they may be useful if a dynamical climate model is not skilful for a particular region. Impact surfaces can allow a decision maker to visualise uncertainty in a forecast and have the advantage of summarizing complex modelling in a concise easy to understand way, especially if they are extended to include other disease models and a more extensive quantification of malaria model uncertainty.

Malaria has been the focus of this thesis, however the methods can easily be applied to another disease, or another climate impact. If there is a clear link to climate and a model can be created which using climate model output as input, the flow chart in figure 9.1 applies: climate model validation should always precede impact model validation, tier-3 validation should precede tier-2, and when only tier-2 validation is possible uncertainty is larger. When there is no skill at any validation level then our confidence in forecasts is

lowest and uncertainty is correspondingly highest.

9.3 Limitations and extensions to the research

There are certain caveats associated with this work. Firstly, upper and lower tercile categories have been used throughout as proxy for high risk and low risk. It may be the case that tercile categories are too broad for certain users; perhaps quartiles or quintiles are more appropriate in certain situations. The definition of an event is dependent on the context of the decision to be made and should be discussed specifically with users.

Secondly there are caveats relating to seasonal targets. For instance, skill has been assessed in climate models by looking at three month seasonal averages of temperature and precipitation. In the cases where scores are low it cannot be concluded that there is no skill, or that models cannot forecast here. Even if no evidence for good prediction cannot be found, it may be such that other targets are more skilful (for example by looking at August-October averages instead of July-September). Defining seasons to maximise skill applies equally to the definition of the malaria season.

Furthermore, only skill in temperature and precipitation have been considered here. There may be skill in other variables. For instance in regions where model precipitation does not have skill, variables relating to circulation patterns such as 850hPa geopotential height or surface pressure may have skill and could improve predictions of seasonal rainfall totals (through the use of model output statistics for instance, Glahn and Lowry, 1972).

Finally there is a large uncertainty related to tier-2 validation. As discussed previously, the lack of epidemiological data prevents direct validation of a disease model. Instead a climate-driven disease model can be compared to a disease model driven by reanalysis. The uncertainty here arises from the fact that reanalysis is not exactly representative of reality and that a disease model driven by reanalysis is not exactly representative of disease incidence. To minimise the risk of any false conclusions drawn from this, tier-2 validation should be carried out with multiple reanalysis datasets and the output from the reanalysis-driven model should be compared with real disease data, where this exists.

When considering extensions to this work, research could progress in several directions. For the decadal validation of chapter 4, skill of alternative variables could be considered. It may also be possible to predict climate impacts at decadal timescales by using dynamical or statistical methods focusing only on the evolution of low frequency oceanic oscillations such as the Atlantic Multidecadal Oscillation (AMO) and the Pacific

Decadal Oscillation and then relating them to climate impacts (for example, using the method for predicting AMO shifts in Enfield and Cid-Serrano, 2006). Finally results should eventually be compared to the skill of the next generation of decadal models produced as part of CMIP5, looking for any improvement.

For seasonal climate prediction an obvious extension is to look at a larger number of subregions. Regions used in this thesis were chosen on an *ad hoc* basis; a full exploration of seasonal prediction skill would involve a complete study of all possible regions where diseases have climate links. Other variables could also be considered, and forecasts from the dynamical System 4 could be compared with the best potential forecasts from statistical climate models.

Work done with System 4 and the LMM should be extended by comparing results with calibrated System 4 runs produced at the ECMWF, allowing an evaluation of the effect of precipitation biases on the output of the LMM. As described above, alternative reanalysis could also be used for tier-2 validation, and as a priority a concerted effort should be made to get as much data for tier-3 validation as possible. Skill of the LMM driven by System 4 should also be compared with skill of alternative malaria models driven by System 4, such as VECTRI (Tompkins and Ermert, 2013).

Finally, it would be interesting to repeat the work contained in chapter 8 using synthetic time series from by a weather generator (rather than using daily reanalysis data). This would enable the creation of as large a sample of input data as needed, allowing a fuller exploration of the behaviour of the LMM. Using the methods of this chapter, there is also a clear path to a fuller investigation of malaria model uncertainty, by varying parameters to investigate their uncertainty and incorporating alternative malaria models in a multi-disease model approach. This will advance work toward the goal of fully quantifying the uncertainty in climate-driven disease prediction.

9.4 Final thoughts

This thesis has focused on prediction of climate-driven disease risk at seasonal and decadal timescales. One timescale is which is absent is that of climate change; the predictions for what will happen at the end of the 21st century. Model predictions at this timescale might be considered less useful than those at shorter timescales, one reason being the invalidity of the stationarity assumption.

Stationarity assumes the future will be like the past. Validating against the past to give confidence in a future prediction assumes stationarity, and this has variable validity. For instance, if weather prediction is the goal and a model has been working fairly well

every day for decade, it is safe to assume stationarity and the system has not changed significantly.

The stationarity assumption is less valid the further into the future a prediction is made. For decadal and longer predictions where multi-decadal and centenary oscillations could change the state, and for climate change scenarios, it is least valid. The nature of the system is more likely to change the longer ahead a prediction is made: fundamental dynamics of the ocean-atmosphere system could be modified (e.g. monsoon dynamics) or non-linear changes could occur, rendering projections inaccurate (e.g. permafrost melting). Whilst global average temperature predictions, and those relating to certain aspects of climate (e.g. polar wettening, Mediterranean drying) have more confidence associated with them than others do, regional climate change predictions of temperature and especially rainfall have high uncertainty. Furthermore, the range of potential futures estimated for the end of the century is likely to be an underestimate.

These uncertain projections of regional climate change will propagate through a disease model. There is already an uncertainty in a disease model itself, which increases the bounds of future possibility. Arguably, the invalidity of the stationarity assumption is more relevant for disease than it is for climate. Mosquito-parasite-host dynamics could change (e.g. by adaptation or evolution caused by relatively high replication rates), such that a certain disease could become prevalent in an area where previously the climate was unsuitable. Mosquitoes could adapt to drying regimes, or to warmer climates.

There are also other unknowns to consider: population movement due to urbanisation and disasters (natural and man-made), or the societal response to climate change and technological development. Changes in land use will certainly affect disease dynamics. These forcings could be included in models, however with increasing complexity comes increasing uncertainty; predictions of each forcing come with their own uncertainties, inflating the range of possible futures such that it may increase beyond usefulness.

Useful predictions may be teased out by attempting to predict the effect of one system on another whilst holding everything else constant (or perhaps using a scenario based approach for different futures). As an academic exercise it may be interesting to consider how climate change can impact the disease landscape; research into climate change and disease can also identify emerging diseases and add weight to the (already-overloaded) argument for greenhouse gas mitigation. However it is not obvious that a prediction for malaria in 2080 is useful, if all the quantified and unquantified uncertainties are taken into account. Put another way, what should a decision maker do with the information that there may be a reduction or increase in malaria by the end of the century, when malaria is present now and political and humanitarian funding timescales are short.

After conducting the work presented in this thesis, it is the opinion of the author that for societally beneficial climate-driven disease prediction (and arguably impact prediction in general), seasonal timescales should take priority: at seasonal timescales model validation is possible and forecasts are clearly actionable (e.g. Tall et al., 2012). This focus by impact modellers should be done with an eye on when (and if) predictions at decadal timescales improve.

Ultimately if uncertainties are fully considered and quantified, climate based disease risk forecasts may only be possible a few months ahead of an event. Despite this, there is an understandable demand from decision makers for forecasts at decadal timescales and beyond. To gain their attention and patronage, researchers may downplay the larger uncertainties and offer the impossible. To resist this temptation and maintain integrity, a paraphrased quote from Voltaire should be considered. When it comes to the future:

> *Uncertainty is uncomfortable, but certainty is absurd.*

Bibliography

Adler, R. F., Susskind, J., Huffman, G. J., Bolvin, D., Nelkin, E., Chang, A., Ferraro, R., Gruber, A., Xie, P.-P., Janowiak, J., Rudolf, B., Schneider, U., Curtis, S., and Arkin, P. (2003). "The Version-2 Global Precipitation Climatology Project (GPCP) Monthly Precipitation Analysis (1979 - Present)". *Journal of Hydrometeorology* 4.6, pp. 1147–1167.

Anyamba, A., Linthicum, K. J., and Tucker, C. J. (2001). "Climate-disease connections: Rift Valley Fever in Kenya." *Cadernos de saúde pública / Ministério da Saúde, Fundação Oswaldo Cruz, Escola Nacional de Saúde Pública* 17 Suppl, pp. 133–40.

Baecher, G. B. and Christian, J. T. (2000). "Natural Variation, Limited Knowledge, and the Nature of Uncertainty in Risk Analysis". *'Risk-Based Decisionmaking in Water Resources IX', Oct. 15-20, 2000, Santa Barbara.* 2, pp. 1–16.

Bayoh, M. (2001). "Studies on the development and survival of Anopheles gambiae sensu stricto at various temperatures and relative humidities". PhD thesis. University of Durham (unpublished).

Bhattacharya, S, Sharma, C, Dhiman, R. C., and Mitra, A. P. (2006). "Climate change and malaria in India. (Special section: Climate change and India)". *Current Science* 90, pp. 369–375.

Bouali, L., Philippon, N., Fontaine, B., and Lemond, J. (2008). "Performance of DEMETER calibration for rainfall forecasting purposes: Application to the July-August Sahelian rainfall". *Journal of Geophysical Research* 113.D15, p. D15111.

Box, G. E. P. (1979). "Robustness in the strategy of scientific model building". *Robustness in Statistics*. Ed. by R. L. Launer and G. N. Wilkinson. Academic Press, pp. 201–236.

Bradley, A. A., Schwartz, S. S., and Hashino, T. (2008). "Sampling Uncertainty and Confidence Intervals for the Brier Score and Brier Skill Score". *Weather and Forecasting* 23.5, pp. 992–1006.

Brier, G. W. (1950). "Verification of forecasts expressed in terms of probability". *Monthly Weather Review* 78.1, pp. 1–3.

Bröcker, J. and Smith, L. a. (2007). "Increasing the Reliability of Reliability Diagrams". *Weather and Forecasting* 22.3, pp. 651–661.

Cane, M. A. (2010). "Climate science: Decadal predictions in demand". *Nature Geoscience* 3.4, pp. 231–232.

Collins, M. (2007). "Ensembles and probabilities: a new era in the prediction of climate change". *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences* 365.1857, pp. 1957–1970.

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., Gleason, B. E., Vose, R. S., Rutledge, G., Bessemoulin, P., Brönnimann, S., Brunet, M., Crouthamel, R. I., Grant, a. N., Groisman, P. Y., Jones, P. D., Kruk, M. C., Kruger, a. C., Marshall, G. J., Maugeri, M., Mok, H. Y., Nordli, O., Ross, T. F., Trigo, R. M., Wang, X. L., Woodruff, S. D., and Worley, S. J. (2011). "The Twentieth Century Reanalysis Project". *Quarterly Journal of the Royal Meteorological Society* 137.654, pp. 1–28.

Craig, M. H., Snow, R. W., and Sueur, D le (1999). "A climate-based distribution model of malaria transmission in sub-Saharan Africa." *Parasitology today* 15.3, pp. 105–11.

Dee, D. P., Uppala, S. M., Simmons, a. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. a., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, a. C. M., Berg, L. van de, Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, a. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kå llberg, P., Köhler, M., Matricardi, M., McNally, a. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., Rosnay, P. de, Tavolato, C., Thépaut, J.-N., and Vitart, F. (2011). "The ERA-Interim reanalysis: configuration and performance of the data assimilation system". *Quarterly Journal of the Royal Meteorological Society* 137.656, pp. 553–597.

Descartes, R. (1641). *Meditations on First Philosophy*.

Detinova, T. S. (1962). "Age-grouping methods in Diptera of medical importance with special reference to some vectors of malaria". *Monograph series. World Health Organization* 47, pp. 13–191.

Dool, H. M. V. (2007). *Empirical methods in short-term climate prediction*. Oxford University Press.

Enfield, D. B. and Cid-Serrano, L. (2006). "Projecting the risk of future climate shifts". *International Journal of Climatology* 26.7, pp. 885–895.

Ermert, V., Fink, A. H., Jones, A. E., and Morse, A. P. (2011a). "Development of a new version of the Liverpool Malaria Model. I. Refining the parameter settings and mathematical formulation of basic processes based on a literature review." *Malaria Journal* 10.1, p. 35.

— (2011b). "Development of a new version of the Liverpool Malaria Model. II. Calibration and validation for West Africa." *Malaria Journal* 10.1, p. 62.

Ermert, V., Fink, A. H., Morse, A. P., Jones, A. E., Paeth, H., Di Giuseppe, F., and Tompkins, A. M. (2012). "Development of dynamical weather-disease models to project and forecast malaria in Africa". *Malaria Journal* 11.Suppl 1, P133.

EUROSIP (2013). *ECMWF website*. URL: http://www.ecmwf.int/products/forecasts/seasonal/documentation/eurosip/index.html (Retrieved 11/03/2013).

Frenkel, Y., Majda, A. J., and Khouider, B. (2012). "Using the Stochastic Multicloud Model to Improve Tropical Convective Parameterization: A Paradigm Example". *Journal of the Atmospheric Sciences* 69.3, pp. 1080–1105.

GFDL (2013). *Geophysical Fluid Dynamics Laboratory: GCM Schematic*. URL: http://www.gfdl.noaa.gov/pix/model_development/climate_modeling/climatemodel.png (Retrieved 11/03/2013).

Gigerenzer, G. (2003). "Why does framing influence judgment?" *Journal of General Internal Medicine* 18.11, pp. 960–961.

Gill, C. A. (1923). "The prediction of malaria epidemics". *Indian Journal Of Medical Research* 10, pp. 1136–1143.

Githeko, A. K., Lindsay, S. W., Confalonieri, U. E., and Patz, J. a. (2000). "Climate change and vector-borne diseases: a regional analysis". *Bulletin of the World Health Organization* 78.9, pp. 1136–47.

Glahn, H. R. and Lowry, D. A. (1972). "The Use of Model Output Statistics (MOS) in Objective Weather Forecasting". *Journal of Applied Meteorology* 11.8, pp. 1203–1211.

Gray, V. (2007). "Climate change 2007: the physical science basis summary for policymakers". *Energy & Environment*. Contribution of Working Group I to the Fourth Assessment Report of the IPCC 18.3. Ed. by S. Solomon, D. Qin, M. Manning, Z Chen, M Marquis, K. B. Averyt, M Tignor, and H. L. Miller, pp. 433–440.

Grover-Kopec, E., Kawano, M., Klaver, R. W., Blumenthal, B., Ceccato, P., and Connor, S. J. (2005). "An online operational rainfall-monitoring resource for epidemic malaria early warning systems in Africa." *Malaria Journal* 4, p. 6.

Gubler, D. J., Reiter, P, Ebi, K. L., Yap, W, Nasci, R, and Patz, J. A. (2001). "Climate variability and change in the United States: potential impacts on vector- and rodent-borne diseases". *Environmental Health Perspectives* 109.Suppl 2, pp. 223–233.

Hagedorn, R. and Smith, L. A. (2009). "Communicating the value of probabilistic forecasts with weather roulette". *Meteorological Applications* 16.2, pp. 143–155.

Halpern, J. Y. (2003). *Reasoning about uncertainty*. MIT Press, pp. 1–497.

Hardy, J. L., Meyer, R. P., Presser, S. B., and Milby, M. M. (1990). "Temporal variations in the susceptibility of a semi-isolated population of Culex tarsalis to peroral infection with western equine encephalomyelitis and St. Louis encephalitis viruses". *The American Journal of Tropical Medicine and Hygiene* 42.2, pp. 241–250.

Harper, K., Uccellini, L. W., Morone, L., Kalnay, E., and Carey, K. (2007). "50th Anniversary of Operational Numerical Weather Prediction". *Bulletin of the American Meteorological Society* 88.5, pp. 639–650.

Hewitt, C. D. and Griggs, D. J. (2004). *Ensembles-Based Predictions of Climate Changes and their Impacts*. Tech. rep. 1. UK Met Office.

Hoshen, M. B. and Morse, A. P. (2004). "A weather-driven model of malaria transmission." *Malaria Journal* 3.1, p. 32.

IPCC (2007). *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by S Solomon, D Qin, M Manning, Z Chen, M Marquis, K. B. Averyt, M Tignor, and H. L. Miller. Vol. Geneva. November. Cambridge University Press.

ISI-MIP (2013). *Inter-Sectoral Impact Model Intercomparison Project homepage*. URL: http://www.pik-potsdam.de/research/climate-impacts-and-vulnerabilities/research/rd2-cross-cutting-activities/isi-mip (Retrieved 11/03/2013).

Janssen, P. H. M., Petersen, a. C., Sluijs, J. P. van der, Risbey, J. S., and Ravetz, J. R. (2005). "A guidance for assessing and communicating uncertainties." *Water Science and Technology : a Journal of the International Association on Water Pollution Research* 52.6, pp. 125–31.

Jolliffe, I. T. and Stephenson, D. B., eds. (2003). *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Vol. 22. 2. Wiley, pp. 403 –405.

Jones, A. E. and Morse, A. P. (2012). "Skill of ENSEMBLES seasonal re-forecasts for malaria prediction in West Africa". *Geophysical Research Letters* 39.23, p. L23707.

Jones, A. E. and Morse, A. P. (2010). "Application and Validation of a Seasonal Ensemble Prediction System Using a Dynamic Malaria Model". EN. *Journal of Climate* 23.15, pp. 4202–4215.

Jones, A. E. (2007). "Seasonal ensemble prediction of malaria in Africa". PhD thesis. University of Liverpool (unpublished).

Joslyn, S., Pak, K., Jones, D., Pyles, J., and Hunt, E. (2007). "The Effect of Probabilistic Information on Threshold Forecasts". *Weather and Forecasting* 22.4, pp. 804–812.

Jupp, T. E., Lowe, R., Coelho, C. A. S., and Stephenson, D. B. (2012). "On the visualization, verification and recalibration of ternary probabilistic forecasts." *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences* 370.1962, pp. 1100–20. arXiv:1103.1303.

Kalnay, E, Kanamitsu, M, Kistler, R, Collins, W, Deaven, D, Gandin, L, Iredell, M, Saha, S, White, G, Woollen, J, Zhu, Y, Leetmaa, A, Reynolds, R, Chelliah, M, Ebisuzaki, W, Higgins, W, Janowiak, J, Mo, K. C., Ropelewski, C, Wang, J, Jenne, R., and Joseph, D. (1996).

"The NCEP/NCAR 40-Year Reanalysis Project". *Bulletin of the American Meteorological Society* 77.3, pp. 437–471.

Karoly, D. J. and Wu, Q. (2005). "Detection of Regional Surface Temperature Trends". *Journal of Climate* 18.21, pp. 4337–4343.

Keenlyside, N. S. and Ba, J. (2010). "Prospects for decadal climate prediction". *Wiley Interdisciplinary Reviews: Climate Change* 1.5, pp. 627–635.

Kloprogge, P., Sluijs, J. P. van der, and Wardekker, A. (2007). *Uncertainty Communication: Issues and good practice*. Tech. rep. Copernicus Institute for Sustainable Development and Innovation.

Knol, A. B., Petersen, A. C., Sluijs, J. P. van der, and Lebret, E. (2009). "Dealing with uncertainties in environmental burden of disease assessment." *Environmental Health: A Global Access Science Source* 8, p. 21.

Knutson, T. R., Delworth, T. L., Dixon, K. W., and Stouffer, R. J. (1999). "Model assessment of regional surface temperature trends (1949-1997)". *Journal of Geophysical Research* 104.D24, p. 30981.

Lafferty, K. (2009). "The ecology of climate change and infectious diseases". *Ecology* 90.4, pp. 888–900.

Lindsay, S. W. and Birley, M. H. (1996). "Climate change and malaria transmission." *Annals of Tropical Medicine and Parasitology* 90.6, pp. 573–88.

Lorenz, E. N. (1963). "Deterministic Nonperiodic Flow". *Journal of the Atmospheric Sciences* 20.2, pp. 130–141.

Lynch, P (2008). "The origins of computer weather prediction and climate modeling". *Journal of Computational Physics* 227.7, pp. 3431–3444.

MacLeod, D. A., Caminade, C, and Morse, A. P. (2012). "Useful decadal climate prediction at regional scales? A look at the ENSEMBLES stream 2 decadal hindcasts". *Environmental Research Letters* 7.4, p. 044012.

Mantua, N. and Hare, S. (2002). "The Pacific Decadal Oscillation". *Journal of Oceanography* 58.1, pp. 35–44.

Mantua, N. J., Hare, S. R., Zhang, Y., Wallace, J. M., and Francis, R. C. (1997). "A Pacific Interdecadal Climate Oscillation with Impacts on Salmon Production". *Bulletin of the American Meteorological Society* 78.6, pp. 1069–1079.

Martens, P, Kovats, R. S., Nijhof, S, Vries, P. D., Livermore, M. T. J., Bradley, D. J., Cox, J, and Mcmichael, A. J. (1999). "Climate change and future populations at risk of malaria". *Global Environmental Change* 9, pp. 89–107.

Martens, W., Jetten, T., Rotmans, J, and Niessen, L. (1995). "Climate change and vector-borne diseases". *Global Environmental Change* 5.3, pp. 195–209.

Mason, S. J. and Graham, N. E. (2002). "Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation". *Quarterly Journal of the Royal Meteorological Society* 128.584, pp. 2145–2166.

Mastrandrea, M. D., Field, C. B., Stocker, T. F., Edenhofer, O., Ebi, K. L., Frame, D. J., Held, H., Kriegler, E., Mach, K. J., Matschoss, P. R., Plattner, G.-K., Yohe, G. W., and Zwiers, F. W. (2010). *Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties*. Tech. rep.

McGuffie, K. and Henderson-Sellers, A. (2005). *A Climate Modelling Primer*. Vol. 1. John Wiley & Sons, Ltd, p. 280.

McMichael, A. J., Woodruff, R. E., and Hales, S. (2006). "Climate change and human health: present and future risks." *Lancet* 367.9513, pp. 859–69.

Meehl, G. A., Goddard, L., Murphy, J., Stouffer, R. J., Boer, G., Danabasoglu, G., Dixon, K., Giorgetta, M. a., Greene, A. M., Hawkins, E., Hegerl, G., Karoly, D., Keenlyside, N., Kimoto, M., Kirtman, B., Navarra, A., Pulwarty, R., Smith, D., Stammer, D., and Stockdale, T. (2009). "Decadal Prediction". *Bulletin of the American Meteorological Society* 90.10, pp. 1467–1485.

Mehta, V., Meehl, G., Goddard, L., Knight, J., Kumar, A., Latif, M., Lee, T., Rosati, A., and Stammer, D. (2011). "Decadal Climate Predictability and Prediction: Where Are We?" *Bulletin of the American Meteorological Society* 92.5, pp. 637–640.

Mills, G., Cleugh, H., Emmanuel, R., Endlicher, W., Erell, E., McGranahan, G., Ng, E., Nickson, a., Rosenthal, J., and Steemer, K. (2010). "Climate Information for Improved Planning and Management of Mega Cities (Needs Perspective)". *Procedia Environmental Sciences* 1, pp. 228–246.

Min, S.-K., Zhang, X., and Zwiers, F. (2008). "Human-induced Arctic moistening." *Science (New York, N.Y.)* 320.5875, pp. 518–20.

Mondet, B., Diaïté, A., Ndione, J.-A., Fall, A. G., Chevalier, V., Lancelot, R., Ndiaye, M., and Ponçon, N. (2005). "Rainfall patterns and population dynamics of Aedes (Aedimorphus) vexans arabiensis, Patton 1905 (Diptera: Culicidae), a potential vector of Rift Valley Fever virus in Senegal." *Journal of Vector Ecology : Journal of the Society for Vector Ecology* 30.1, pp. 102–6.

Morse, A., Doblas-Reyes, F. J., Hoshen, M., Hagedorn, R, and Palmer, T. N. (2005). "A forecast quality assessment of an end-to-end probabilistic multi-model seasonal forecast system". *Tellus*, pp. 464–475.

Murphy, A. H. (1969). "On Expected-Utility Measures in Cost-Loss Ratio Decision Situations". *Journal of Applied Meteorology* 8.6, pp. 989–991.

Murphy, J., Kattsov, V., Keenlyside, N., Kimoto, M., Meehl, G., Mehta, V., Pohlmann, H., Scaife, A., and Smith, D. (2010). "Towards Prediction of Decadal Climate Variability and Change". *Procedia Environmental Sciences* 1, pp. 287–304.

Murphy, J. M., Booth, B. B. B., Collins, M, Harris, G. R., Sexton, D. M. H., and Webb, M. J. (2007). "A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles". *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences* 365.1857, pp. 1993–2028.

Najm, S. M. (1966). "The Place and Function of Doubt in the Philosophies of Descartes and Al-Hazali". *Philosophy East and West* 16.3-4, pp. 133–41.

NCAR (2013). *The NCAR Command Language (Version 6.0.0) [Software]*. URL: `http://www.ncl.ucar.edu/index.shtml` (Retrieved 11/03/2013).

Oldenborgh, G. J., Doblas-Reyes, F. J., Wouters, B., and Hazeleger, W. (2012). "Decadal prediction skill in a multi-model ensemble". *Climate Dynamics* 38.7-8, pp. 1263–1280.

Oreskes, N., Stainforth, D. A., and Smith, L. A. (2010). "Adaptation to Global Warming: Do Climate Models Tell Us What We Need to Know?" *Philosophy of Science* 77.5, pp. 1012–1028.

Palmer, T. N. (2000). "Predicting uncertainty in forecasts of weather and climate". *Reports on Progress in Physics* 63.2, pp. 71–116.

Palmer, T. N., Alessandri, A, Andersen, U, Cantelaube, P, Davey, M, Delecluse, P, Deque, M, Diez, E, Doblas-Reyes, F. J., Feddersen, H, Graham, R, Gualdi, S, Gueremy, J.-F., Hagedorn, R, Hoshen, M, Keenlyside, N, Latif, M, Lazar, A, Maisonnave, E, Marletto, V, Morse, A. P., Orfila, B, Rogel, P, Terres, J.-M., and Thomson, M. C. (2004). "Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER)". *Bulletin of the American Meteorological Society* 85.6, pp. 853–872.

Palmer, T. N. and Hagedorn, R (2006a). *Predictability of weather and climate: from theory to practice*. Ed. by T. Palmer and R. Hagedorn. Cambridge University Press. Chap. 1.

Palmer, T. N. and Williams, P. D. (2008). "Introduction. Stochastic physics and climate modelling." *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences* 366.1875, pp. 2421–7.

Palmer, T. and Hagedorn, R., eds. (2006b). *The Predictability of Weather and Climate*. Vol. 11. 1-2. Cambridge University Press, pp. 239–246.

Pascual, M, Ahumada, J. A., Chaves, L. F., Rodó, X, and Bouma, M (2006). "Malaria resurgence in the East African highlands: Temperature trends revisited". *Proceedings of the National Academy of Sciences of the United States of America* 103.15, pp. 5829–5834.

Pereira, A. G. a. and Quintana, S. C. (2002). "From Technocratic to Participatory Decision Support Systems: Responding to the New Governance Initiatives". *Journal of Geographic Information and Decision Analysis* 6.2, pp. 95–107.

Peterson, A. T. (2003). "Predicting the geography of species' invasions via ecological niche modeling." *The Quarterly Review of Biology* 78.4, pp. 419–33.

Peterson, A. and Shaw, J. (2003). "Lutzomyia vectors for cutaneous leishmaniasis in Southern Brazil: ecological niche models, predicted geographic distributions, and climate change effects". *International Journal for Parasitology* 33.9, pp. 919–931.

Pohlmann, H., Jungclaus, J. H., Köhl, A., Stammer, D., and Marotzke, J. (2009). "Initializing Decadal Climate Predictions with the GECCO Oceanic Synthesis: Effects on the North Atlantic". *Journal of Climate* 22.14, pp. 3926–3938.

QWeCI Project (2011). *D2.1a Report on dynamic malaria model runs for regions of interest and verification against datasets*. Tech. rep.

Reeves, W. C., Hardy, J. L., Reisen, W. K., and Milby, M. M. (1994). "Potential effect of global warming on mosquito-borne arboviruses". *Journal of Medical Entomology* 31.3, pp. 323–32.

Reisen, W. K., Lothrop, H. D., and Hardy, J. L. (1995). "Bionomics of Culex tarsalis (Diptera: Culicidae) in relation to arbovirus transmission in southeastern California." *Journal of Medical Entomology* 32.3, pp. 316–27.

Reisen, W. K. (2010). "Landscape epidemiology of vector-borne diseases". *Annual Review of Entomology* 55, pp. 461–83.

Reiter, P (2001). "Climate change and mosquito-borne disease". *Environmental Health Perspectives* 109 Suppl.September 2000, pp. 141–61.

Richardson, C. W., Barrow, E. M., Semenov, M. A., and Brooks, R. J. (1998). "Comparison of the WGEN and LARS-WG stochastic weather generators for diverse climates". *Climate Research* 10.2, pp. 95–107.

Richardson, D. S. (2000). "Skill and relative economic value of the ECMWF ensemble prediction system". *Quarterly Journal of the Royal Meteorological Society* 126.563, pp. 649–667.

Rodó, X., Ballester, J., Cayan, D., Melish, M. E., Nakamura, Y., Uehara, R., and Burns, J. C. (2011). "Association of Kawasaki disease with tropospheric wind patterns." *Scientific reports* 1.152.

Ruti, P. M., Williams, J. E., Hourdin, F., Guichard, F., Boone, a., Van Velthoven, P., Favot, F., Musat, I., Rummukainen, M., Domínguez, M., Gaertner, M. A., Lafore, J. P., Losada, T., Rodriguez de Fonseca, M. B., Polcher, J., Giorgi, F., Xue, Y., Bouarar, I., Law, K., Josse, B., Barret, B., Yang, X., Mari, C., and Traore, a. K. (2011). "The West African climate system: a review of the AMMA model inter-comparison initiatives". *Atmospheric Science Letters* 12.1, pp. 116–122.

S4 (2013). *ECMWF website*. URL: http://www.ecmwf.int/products/changes/system4/technical_description.html (Retrieved 11/03/2013).

Schlesinger, M. E. and Ramankutty, N. (1994). "An oscillation in the global climate system of period 65 - 70 years". *Nature* 367.6465, pp. 723–726.

Shea, K. M., Truckner, R. T., Weber, R. W., and Peden, D. B. (2008). "Climate change and allergic disease". *The Journal of Allergy and Clinical Immunology* 122.3, 443–53; quiz 454–5.

Smith, D. M., Eade, R., Dunstone, N. J., Fereday, D., Murphy, J. M., Pohlmann, H., and Scaife, A. (2010). "Skilful multi-year predictions of Atlantic hurricane frequency". *Nature Geoscience* 3.12, pp. 846–849.

Snow, R. W., Craig, M, Deichmann, U, and Marsh, K (1999). "Estimating mortality, morbidity and disability due to malaria among Africa's non-pregnant population." *Bulletin of the World Health Organization* 77.8, pp. 624–40.

Stainforth, D. A., Allen, M. R., Tredger, E. R., and Smith, L. A. (2007). "Confidence, uncertainty and decision-support relevance in climate predictions". 365, pp. 2145–2161.

Stensrud, D. J. (2009). *Parameterization Schemes: Keys to Understanding Numerical Weather Prediction*. Cambridge University Press.

Storch, H. von and Zwiers, F. W. (1984). *Statistical Analysis in Climate Research*. Vol. 95. 452. Cambridge: Cambridge University Press. Chap. 484, p. 1375.

Sultan, B., Labadi, K., Guégan, J.-F., and Janicot, S. (2005). "Climate drives the meningitis epidemics onset in west Africa." *PLoS Medicine* 2.1, e6.

Swaroop, S (1949). "Forecasting of epidemic malaria in the Punjab, India." *The American Journal of Tropical Medicine and Hygiene* 29.1, pp. 1–17.

Tall, A., Mason, S. J., Aalst, M. van, Suarez, P., Ait-Chellouche, Y., Diallo, A. a., and Braman, L. (2012). "Using Seasonal Climate Forecasts to Guide Disaster Management: The Red Cross Experience during the 2008 West Africa Floods". *International Journal of Geophysics* 2012, pp. 1–12.

Tanser, F. C., Sharp, B., and Sueur, D. le (2003). "Potential effect of climate change on malaria transmission in Africa." *Lancet* 362.9398, pp. 1792–8.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A. (2012). "An Overview of CMIP5 and the Experiment Design". *Bulletin of the American Meteorological Society* 93.4, pp. 485–498.

Tennekes, H. (1992). "Karl Popper and the accountability of numerical weather forecasting". *Weather* 47.9, pp. 343–346.

Thomas, C. D., Cameron, A, Green, R. E., Bakkenes, M, Beaumont, L. J., Collingham, Y. C., Erasmus, B. F. N., De Siqueira, M. F., Grainger, A, Hannah, L, Hughes, L, Huntley, B, Van

Jaarsveld, A. S., Midgley, G. F., Miles, L, Ortega-Huerta, M. A., Peterson, A. T., Phillips, O. L., and Williams, S. E. (2004). "Extinction risk from climate change". *Nature* 427.6970, pp. 145–148.

Thomson, M. C., Doblas-Reyes, F. J., Mason, S. J., Hagedorn, R, Connor, S. J., Phindela, T, Morse, A. P., and Palmer, T. N. (2006). "Malaria early warnings based on seasonal climate forecasts from multi-model ensembles". *Nature* 439.7076, pp. 576–579.

Thomson, M. C., Mason, S. J., Phindela, T., and Connor, S. J. (2005). "Use of rainfall and sea surface temperature monitoring for malaria early warning in Botswana." *The American Journal of Tropical Medicine and Hygiene* 73.1, pp. 214–21.

Tompkins, A. M. and Ermert, V. (2013). "A regional-scale, high resolution dynamical malaria model that accounts for population density, climate and surface hydrology." *Malaria journal* 12.1, p. 65.

UKMO (2013). *UK Met Office homepage*. URL: `http://www.metoffice.gov.uk/` (Retrieved 11/03/2013).

Uppala, S. M., K{\AA}llberg, P. W., Simmons, A. J., Andrae, U, Bechtold, V. D. C., Fiorino, M, Gibson, J. K., Haseler, J, Hernandez, A, Kelly, G. A., Li, X, Onogi, K, Saarinen, S, Sokka, N, Allan, R. P., Andersson, E, Arpe, K, Balmaseda, M. A., Beljaars, A. C. M., Berg, L. V. D., Bidlot, J, Bormann, N, Caires, S, Chevallier, F, Dethof, A, Dragosavac, M, Fisher, M, Fuentes, M, Hagemann, S, Hólm, E, Hoskins, B. J., Isaksen, L, Janssen, P. A. E. M., Jenne, R, Mcnally, A. P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N. A., Saunders, R. W., Simon, P, Sterl, A, Trenberth, K. E., Untch, A, Vasiljevic, D, Viterbo, P, and Woollen, J (2005). "The ERA-40 re-analysis". *Quarterly Journal of the Royal Meteorological Society* 131.612, pp. 2961–3012.

Urashima, M., Shindo, N., and Okabe, N. (2003). "A seasonal model to simulate influenza oscillation in Tokyo." *Japanese Journal of Infectious Diseases* 56.2, pp. 43–7.

Van Der Linden, P. and Mitchell, J. F. B. (2009). *ENSEMBLES: Climate Change and its Impacts: Summary of research and results from the ENSEMBLES project*. Tech. rep., p. 160.

Vellinga, M. and Wood, R. A. (2007). "Impacts of thermohaline circulation shutdown in the twenty-first century". *Climatic Change* 91.1-2, pp. 43–63.

Walker, W., Harremoës, P., Rotmans, J, Sluijs, J. van der, Asselt, M. van, Janssen, P, and Krayer von Krauss, M. (2003). "Defining Uncertainty: A Conceptual Basis for Uncertainty Management in Model-Based Decision Support". *Integrated Assessment* 4.1, pp. 5–17.

Wang, B., Zou, X., and Zhu, J. (2000). "Data assimilation and its applications". *Proceedings of the National Academy of Sciences of the United States of America* 97.21, pp. 11143–11144.

Washington, R., Kay, G., Harrison, M., Conway, D., Black, E., Challinor, A., Grimes, D., Jones, R., Morse, A., and Todd, M. (2006). "African Climate Change: Taking the Shorter Route". *Bulletin of the American Meteorological Society* 87.10, pp. 1355–1366.

WCRP (2013). *World Climate Research Programme website*. URL: http://www.wcrp-climate.org/decadal/cmip5.shtml (Retrieved 11/03/2013).

Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences*. 3rd. International Geophysics. Academic Press, pp. 412–500.

Zhang, X., Zwiers, F. W., Hegerl, G. C., Lambert, F. H., Gillett, N. P., Solomon, S., Stott, P. a., and Nozawa, T. (2007). "Detection of human influence on twentieth-century precipitation trends." *Nature* 448.7152, pp. 461–5.

Zhu, Y, Toth, Z, Wobus, R, Richardson, D, and Mylne, K. R. (2002). "The economic value of ensemble-based weather forecasts". *Bulletin of the American Meteorological Society* 83, pp. 73–83.

Appendices

APPENDIX A

Extra figures for Chapter 4

Contained in this chapter are extra figures for chapter 4. Maps of climatologies for observed and model datasets, along with correlation maps and trend correlation matrices using alternative reference datasets (NCEP and ERA40 as reference for temperature, and using NCEP, ERA40 and GPCP as references for precipitation). They have been separated from the main chapter for brevity. Results are discussed in the main chapter.

A.1 Climatologies

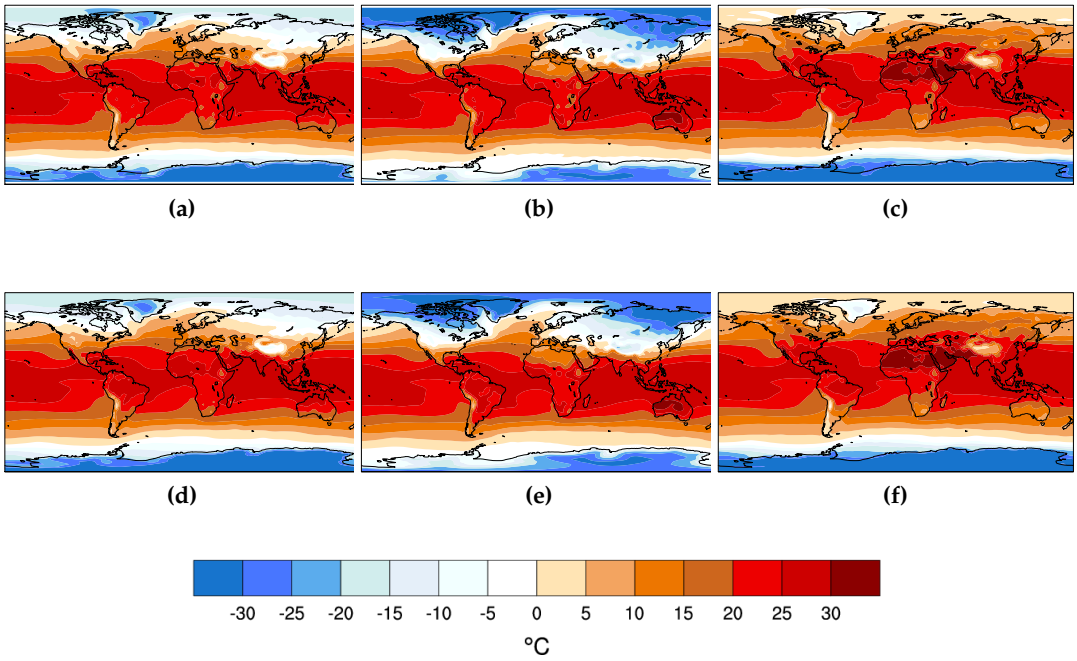


Figure A.1: Observed temperature climatology for 1960-2005. For annual, DJF and JJA averages (left, middle and right column), using the NCEP and ERAINTERIM datasets (top and bottom rows).

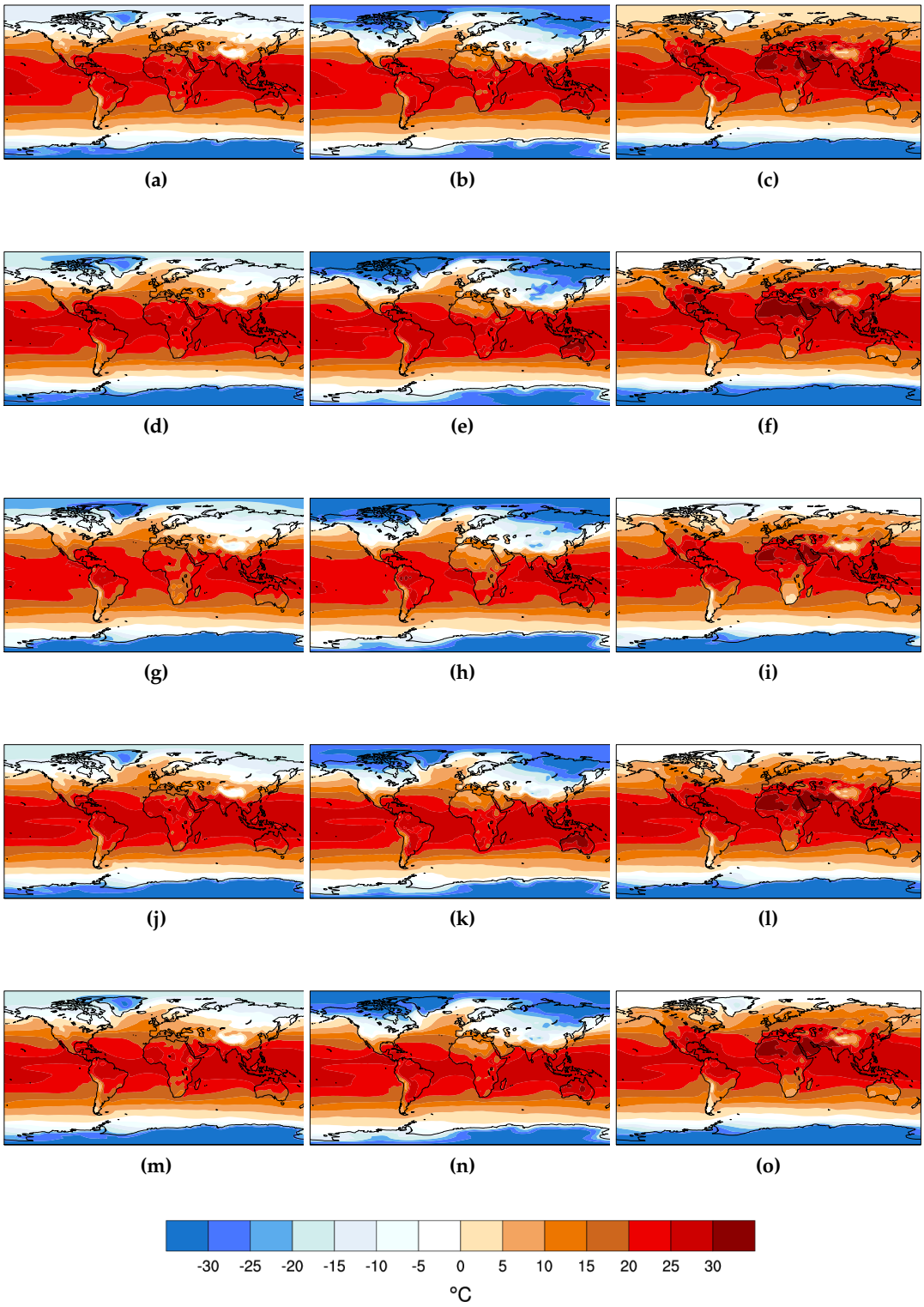


Figure A.2: Annual, DJF and JJA (left, middle and right columns) temperature climatologies from ENSEMBLES stream 2 decadal models. For ECMWF (a-c), UKMO (d-f), CERFACS (g-i), IFM (j-l) and the multimodel mean (m-o)).

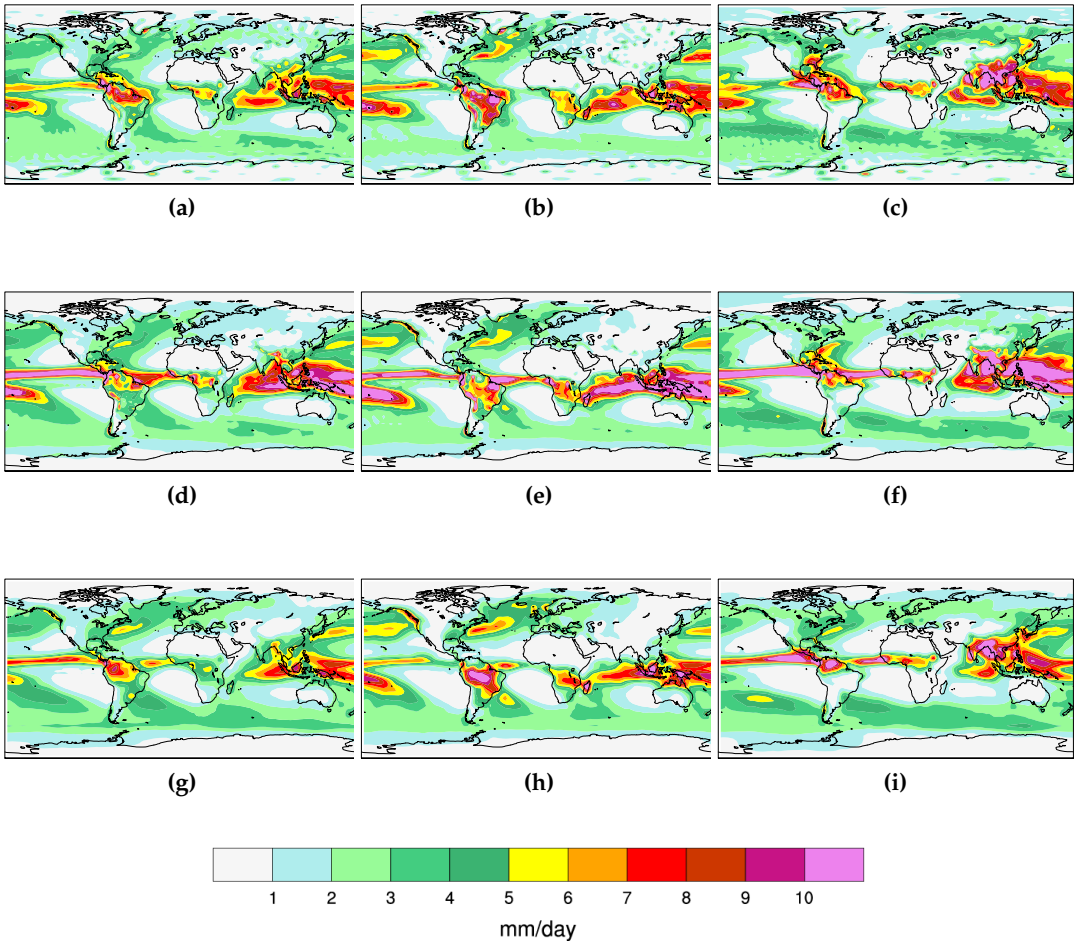


Figure A.3: Observed precipitation climatology of reference datasets. For annual, DJF and JJA averages (left, middle and right columns), using the NCEP, ERA40 and GPCP datasets (top, middle and bottom rows).

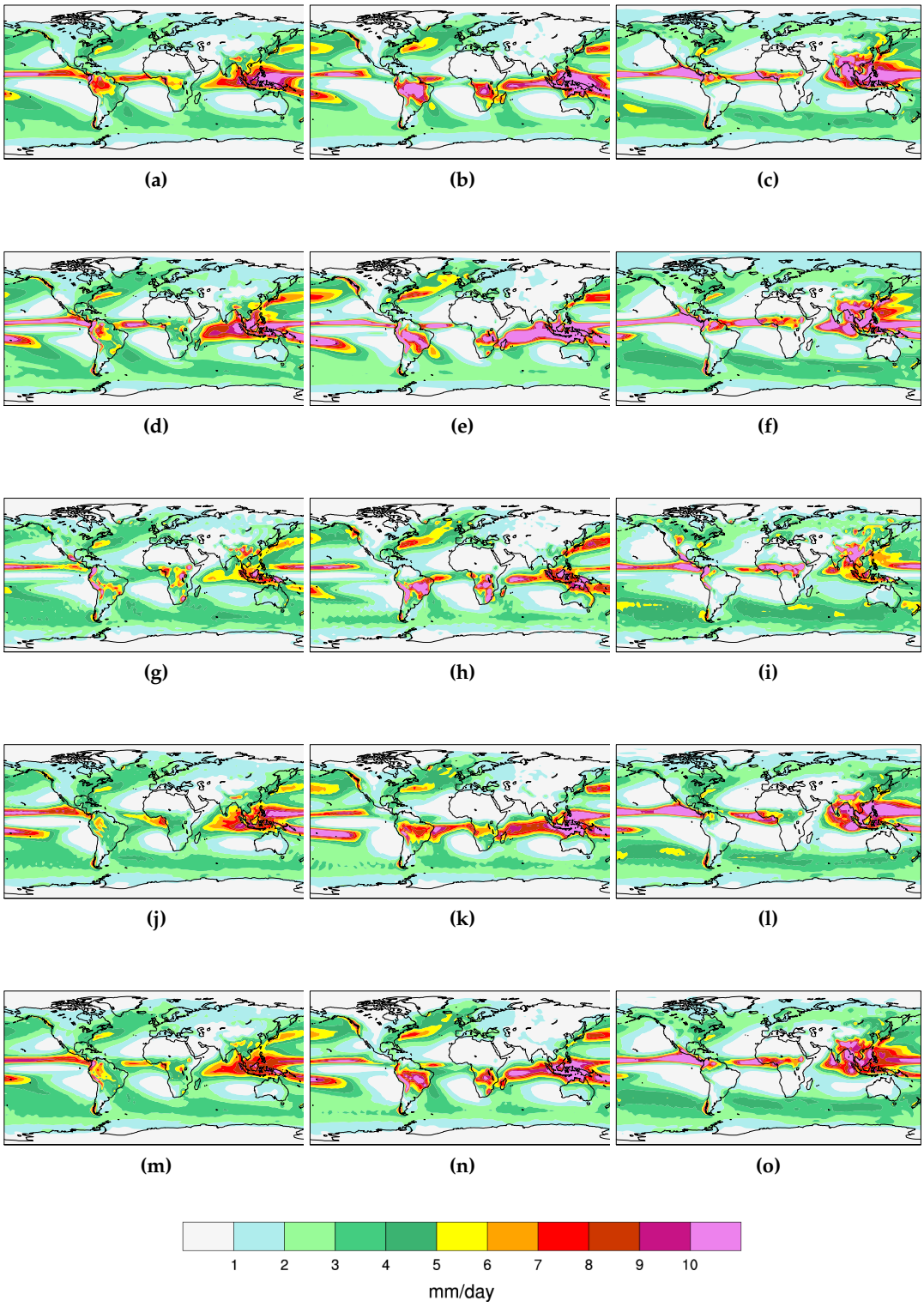


Figure A.4: Annual, DJF and JJA (left, middle and right columns) precipitation climatologies from ENSEMBLES stream 2 decadal models. For ECMWF (a-c), UKMO (d-f), CERFACS (g-i), IFM (j-l) and the multimodel mean (m-o).

A.2 Correlations

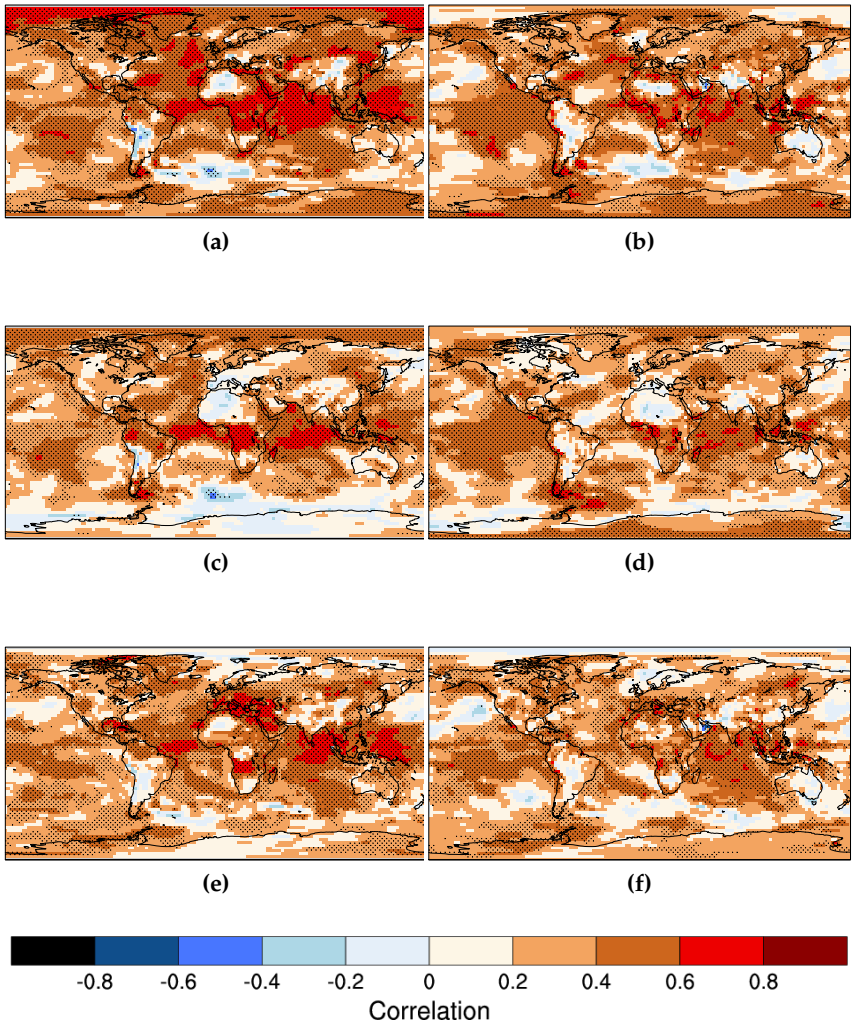


Figure A.5: Temperature correlations between the ensemble mean of the ENSEMBLES decadal hindcasts and NCEP (ERA40) reanalysis shown in the left (right) column. Correlations are shown for annual (a & b), DJF (c & d) and JJA (e & f) averages. Stippled areas indicate significant correlations at the 95% level.

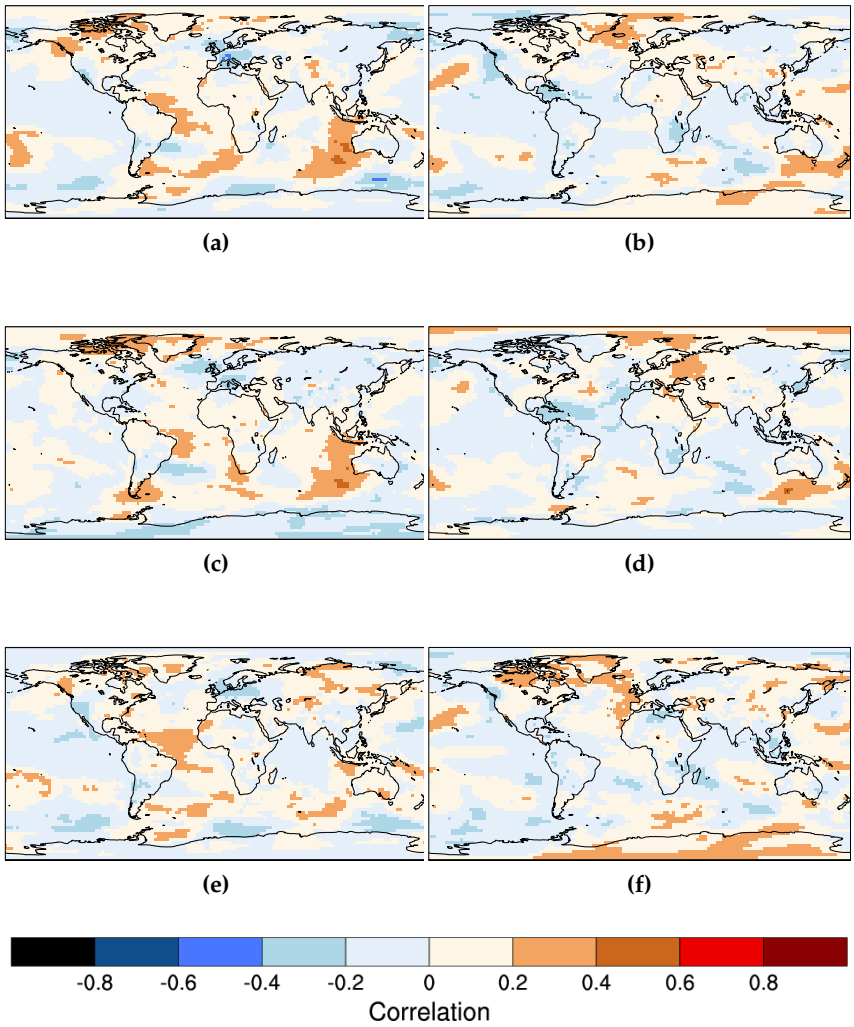


Figure A.6: Temperature correlations between the detrended ensemble mean of the ENSEMBLES decadal hindcasts and NCEP (ERA40) reanalysis shown in the left (right) column. Correlations are shown for annual (a & b), DJF (c & d) and JJA (e & f) averages. Stippled areas indicate significant correlations at the 95% level.

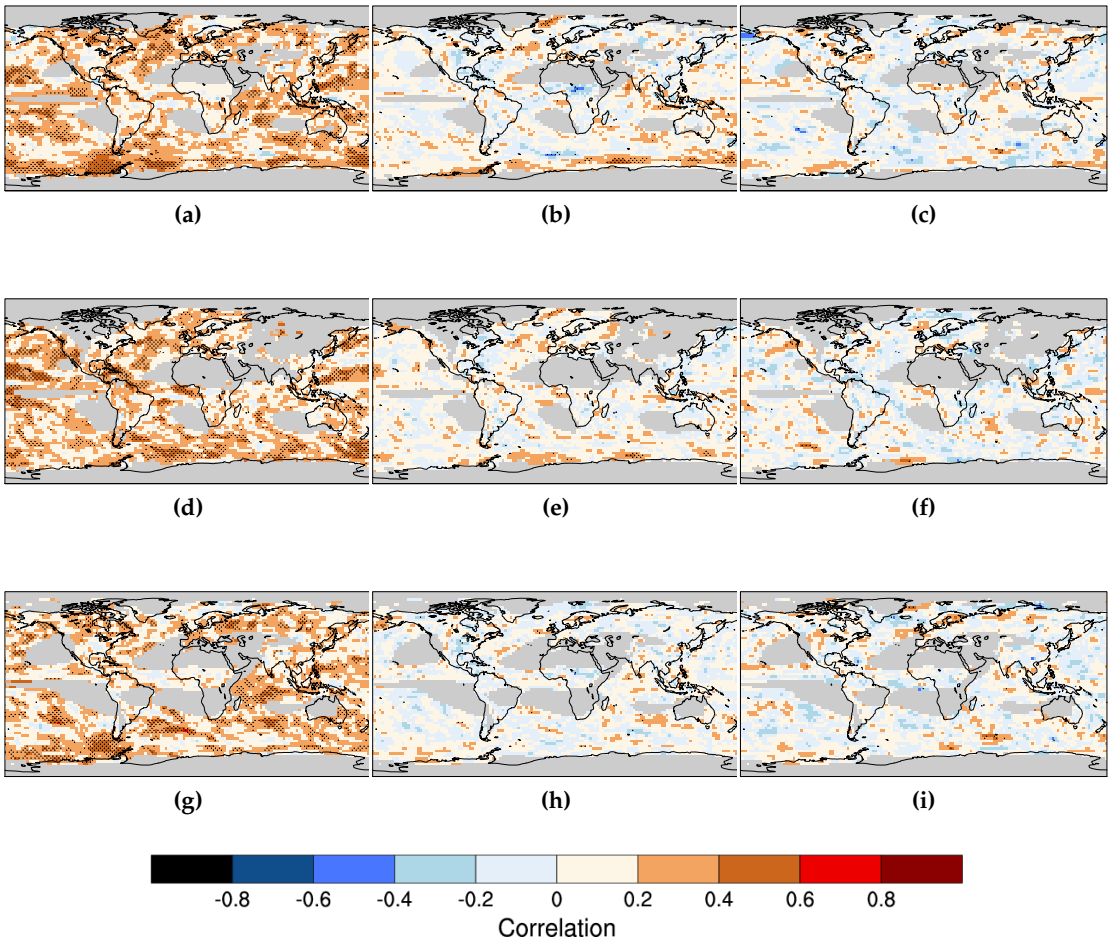


Figure A.7: Precipitation correlations between the ensemble mean of the ENSEMBLES decadal hindcasts and NCEP, ERA40 and GPCP (left, middle, right columns). Correlations are shown for annual (a & b), DJF (c & d) and JJA (e & f) averages. The greyed out area indicates regions where model climatology is less than 1mm/day. Stippled areas indicate significant correlations at the 95% level.

A.3 Trend correlations

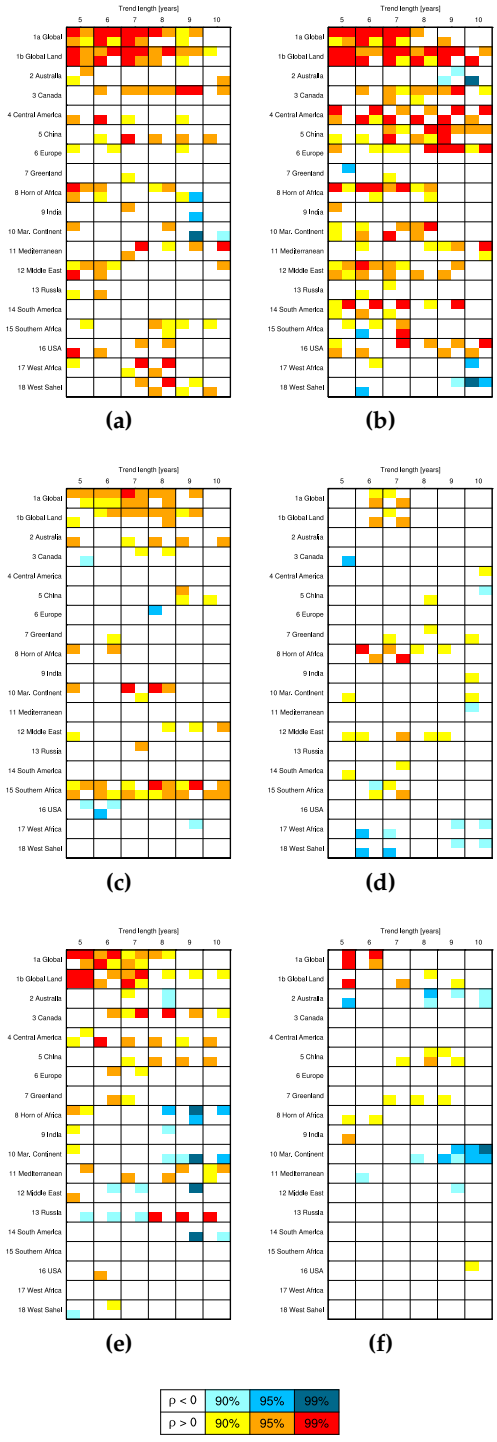


Figure A.8: Multi year trend correlation significance levels for annual, DJF and JJA temperature (top, middle bottom rows). Reference datasets are NCEP and ERA40 (left & right columns). Each quadrant in each square stands for one of the four models in the ENSEMBLES decadal simulations (clockwise from top left: UK Met Office, ECMWF, IFM-GEOMAR, CERFACS). The three variations in warm (cold) colours indicate correlations significantly above (below) zero at the 90%/95%/99% levels respectively (levels at $\pm$0.324, $\pm$0.382, $\pm$0.491 for NCEP and $\pm$0.344, $\pm$0.406,$\pm$0.521 for ERA40).

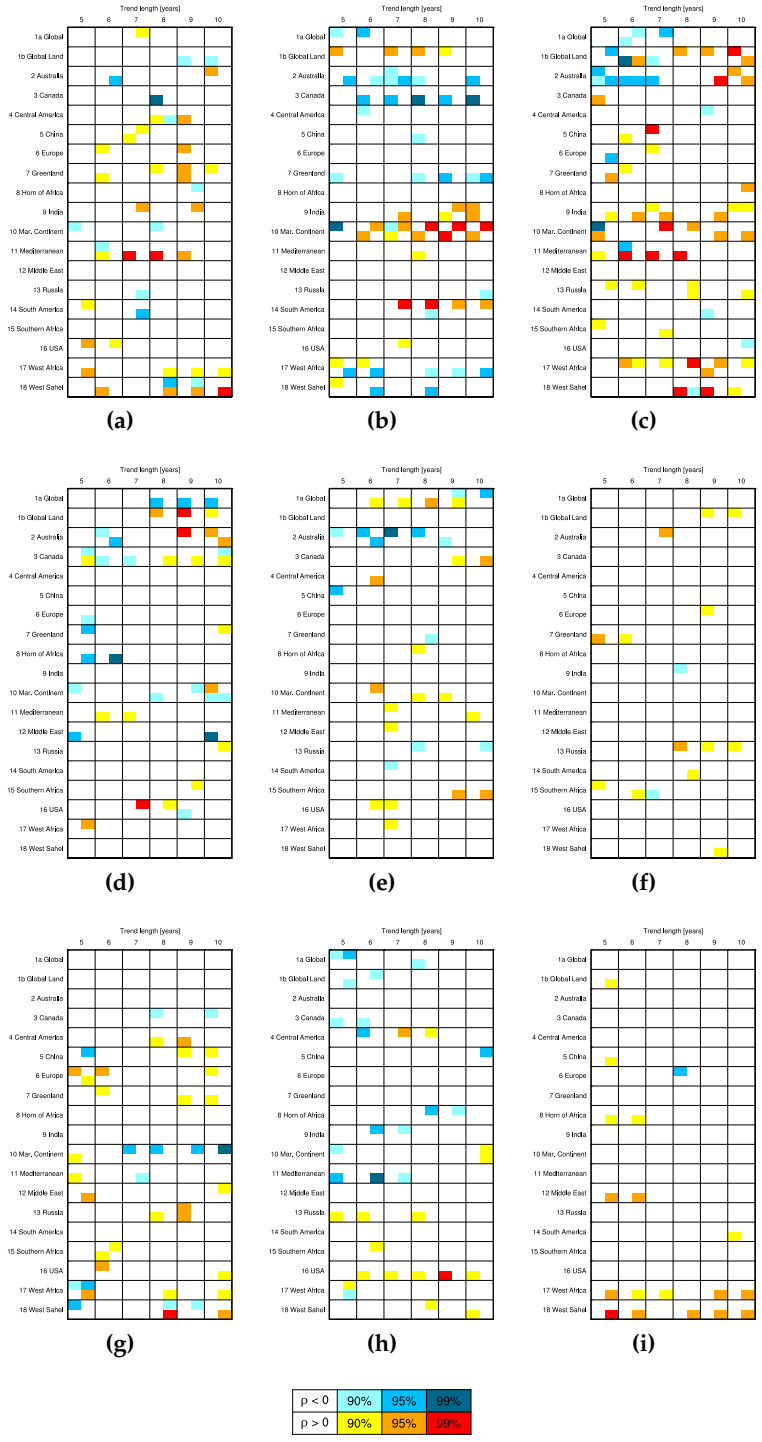


Figure A.9: Multi year trend correlation significance levels for annually averaged precipitation. Reference datasets are NCEP, ERA40 reanalysis and GPCP (left, middle, right column). Each quadrant in each square stands for one of the four models in the ENSEMBLES decadal simulations (clockwise from top left: UK Met Office, ECMWF, IFM-GEOMAR, CERFACS). The three variations in warm (cold) colours indicate correlations significantly above (below) zero at the 90%/95%/99% levels respectively (levels at ± 0.324, ± 0.382, ± 0.491 for NCEP, ± 0.344, ± 0.406,± 0.521 for ERA40 and ± 0.401, ± 0.472, ± 0.600 for GPCP).

APPENDIX B

Extra figures for Chapter 5

Contained in this chapter are extra figures for chapter 5, for East Africa and for the Indian subcontinent.

B.1 East Africa

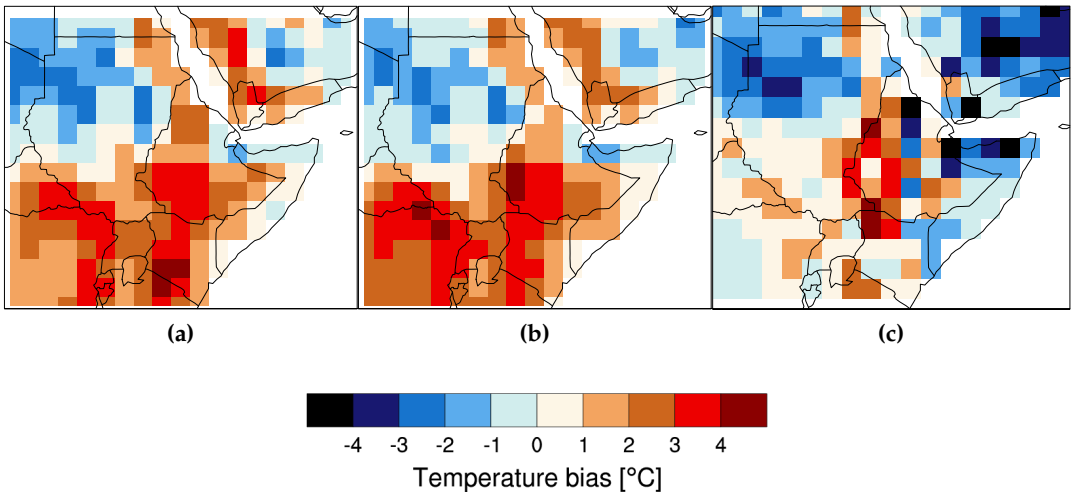


Figure B.1: Ensemble mean temperature bias vs NCEP, for DEMETER, ENSEMBLES and System 4 (a-c).

Biases for temperature for March to May over the Horn of Africa are shown in figure B.1. DEMETER and ENSEMBLES have a similar bias, mostly too warm over the areas where rainfall is present, whilst the temperature is too cold over the desert to the north. There is no significant improvement between the systems, in fact the bias for ENSEMBLES is slightly larger. For System 4 there is a large improvement, with most of the region reduced to a bias of less than one degree.

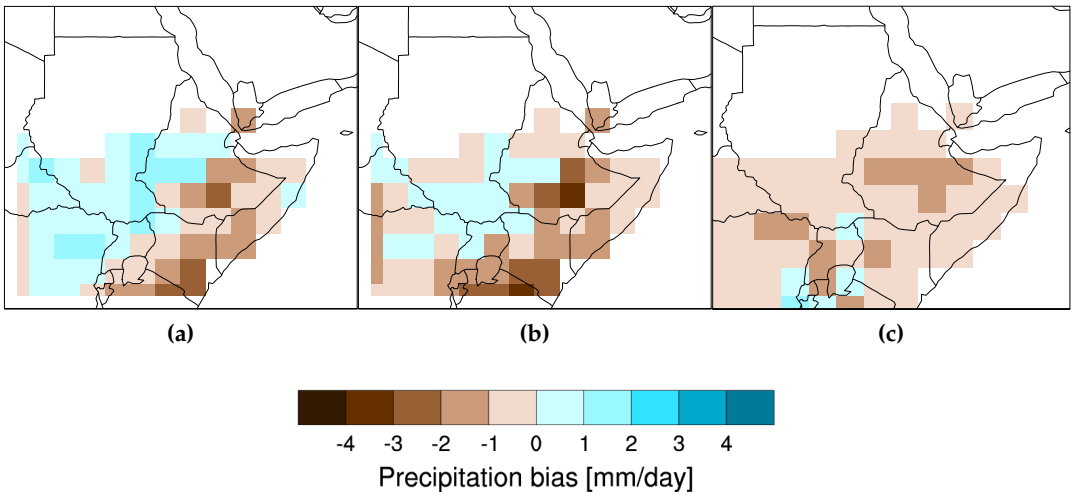


Figure B.2: As figure B.1, for ensemble mean precipitation vs GPCP. Only points are shown where the observed seasonal climatology is greater than 1mm/day.

For precipitation (figure B.2), the bias is again lowest for System 4, with less than mm/day bias over most of the region. For DEMETER and ENSEMBLES the bias is larger; DEMETER has too much rain over South Sudan and not enough rain near the coast to the south, over Kenya. ENSEMBLES has a similar pattern, though the wet bias to the north west is reduced whilst the dry bias is slightly larger.

Correlations for temperature are seen in figure B.3. There is a marked improvement between DEMETER and ENSEMBLES, as can be seen from the much larger extent of correlation above significance. The area of significant correlation is slightly reduced in System 4, being limited to the south west of the region, and the coast around Somalia and Ethiopia.

For precipitation (figure B.4), there is a steady improvement from DEMETER, through ENSEMBLES to System 4. For DEMETER (figure B.4a) the correlation coefficient is low everywhere, mostly under 0.1 and not significant. For ENSEMBLES (figure B.4b) the correlation is higher, though still below significance. System 4 (figure B.4c) has higher values of the correlation coefficient over a wider area, though this is limited to below significance apart from a few gridpoints.

ROC AUC maps are shown in figure B.5 for temperature. There does not appear to be an increase in skill between the systems, and for lower tercile events System 4 has the lowest score. In general the pattern of significant ROC AUC follows that of correlations (see figure B.3). For precipitation ROC AUC (figure B.6) the similarity of the pattern to the correlations is also apparent; DEMETER has ROC AUC mostly around 0.5, with some improvement in ENSEMBLES, where a few gridpoints are significant, mostly for lower

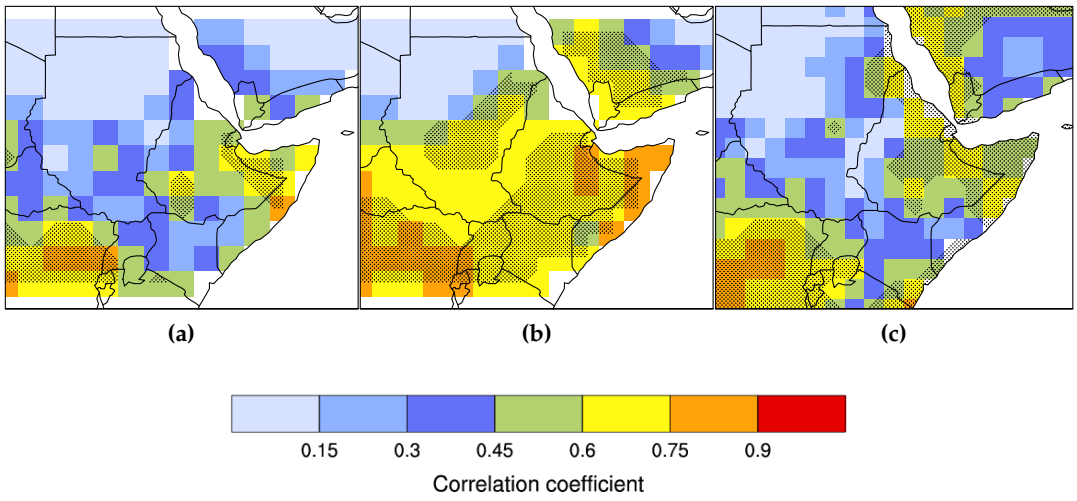


Figure B.3: Pearson's product-moment correlations of MAM ensemble mean precipitation vs NCEP, for DEMETER, ENSEMBLES and System 4 (a-c). Forecasts issued at the start of February. Stippled area shows 99% confidence level, with sample size adjusted to take account for autocorrelation.

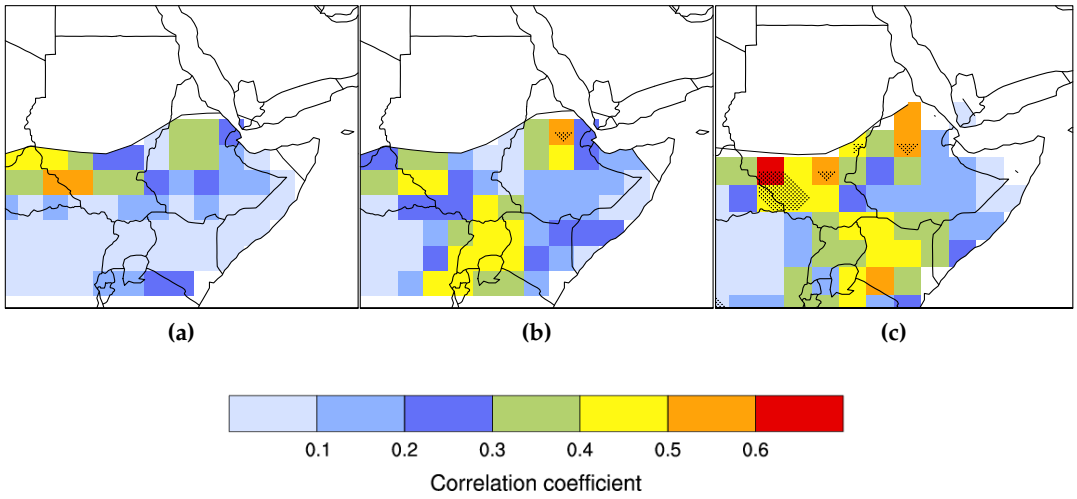


Figure B.4: As figure B.3, for ensemble mean precipitation vs GPCP. Only points are shown where the observed seasonal climatology is greater than 1mm/day.

tercile events. System 4 has higher values of ROC AUC than DEMETER, but they are lower than ENSEMBLES, and there are also fewer points above significance than there are in ENSEMBLES.

For reliability of temperature forecasts over Kenya (figure B.7), it can be seen that there is a steady improvement in reliability between the systems, for upper and for lower tercile forecasts. This is reflected in the BSS for upper tercile forecasts which is highest

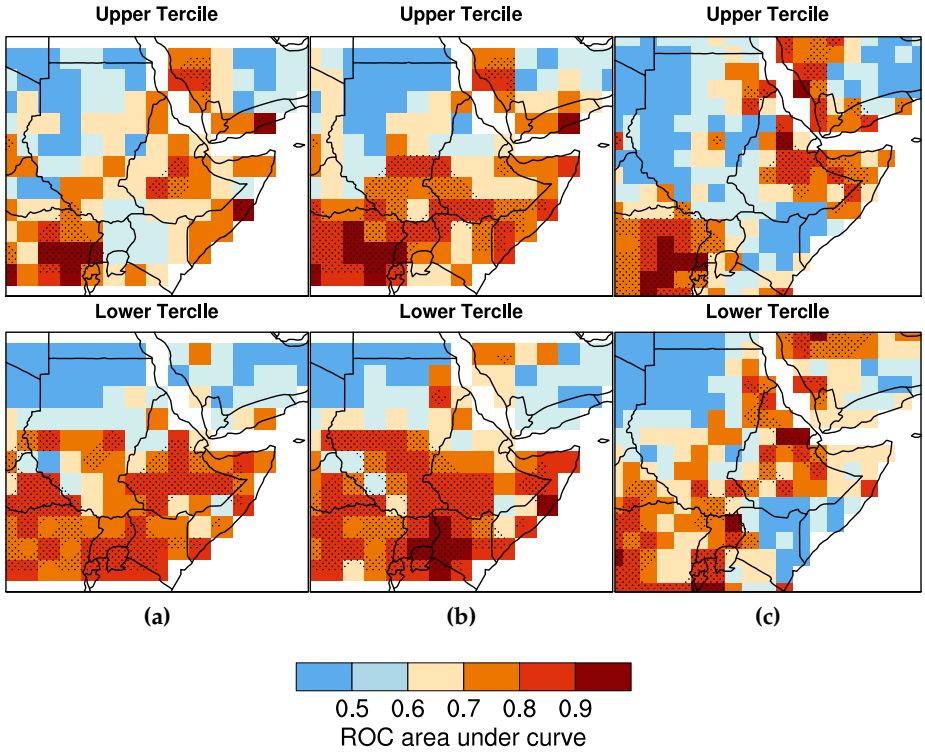


Figure B.5: Relative operating characteristic area under curve (ROC AUC) for MAM temperature vs NCEP, for the February start dates of DEMETER, ENSEMBLES and System 4 (a-c). Stippled area indicates where the AUC is significant at the 95% level.

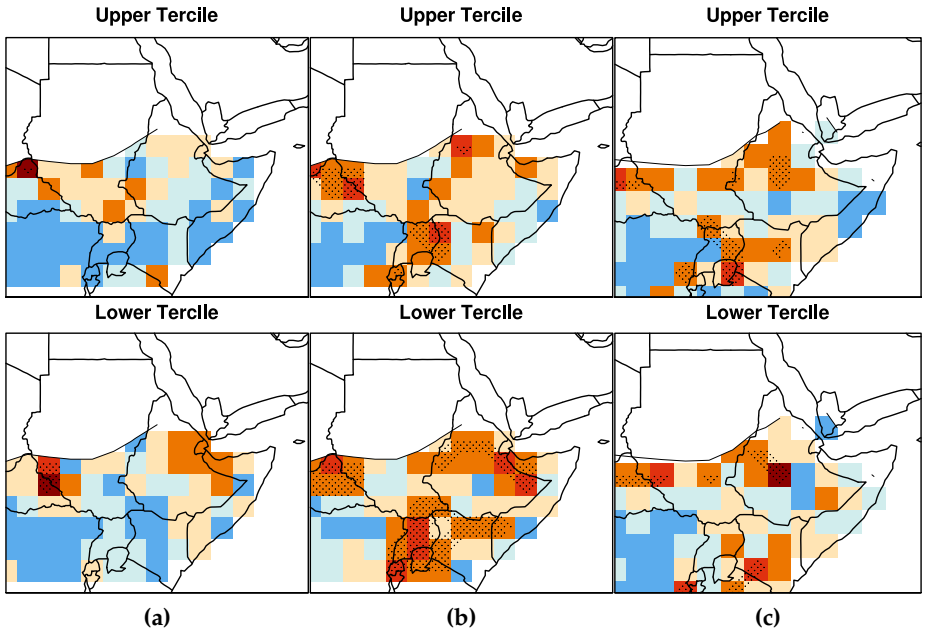


Figure B.6: As figure B.5, for precipitation vs GPCP. Only points are shown where the observed seasonal climatology is greater than 1mm/day.

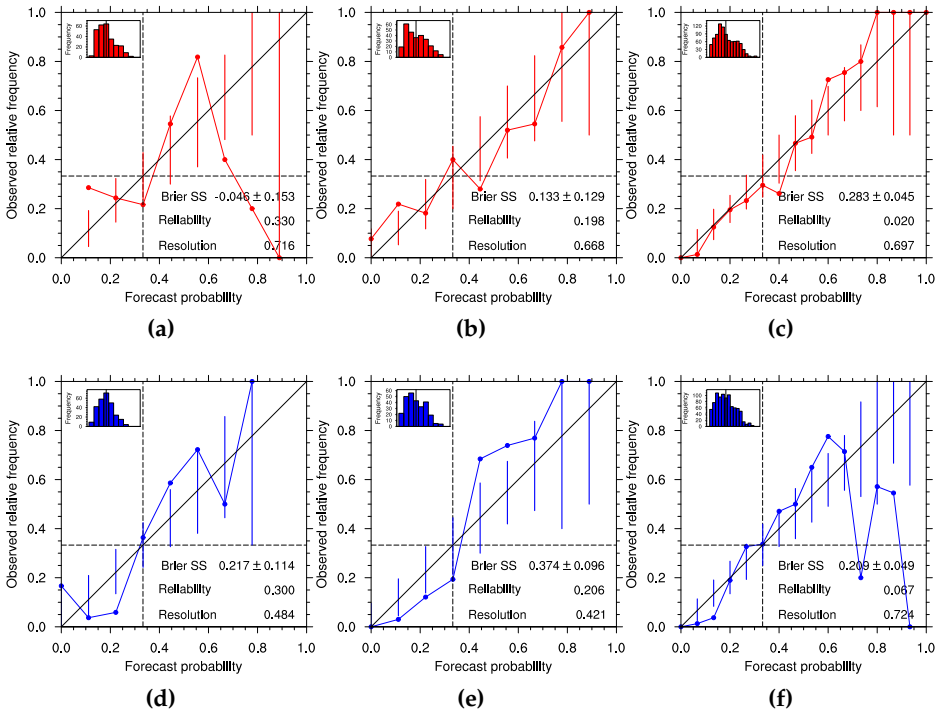


Figure B.7: Reliability of upper (a-c) and lower (d-f) tercile MAM temperature forecasts over Kenya (vs NCEP): DEMETER (a & d), ENSEMBLES (b & e) and System 4 (c & f). Forecasts issued at the start of February.

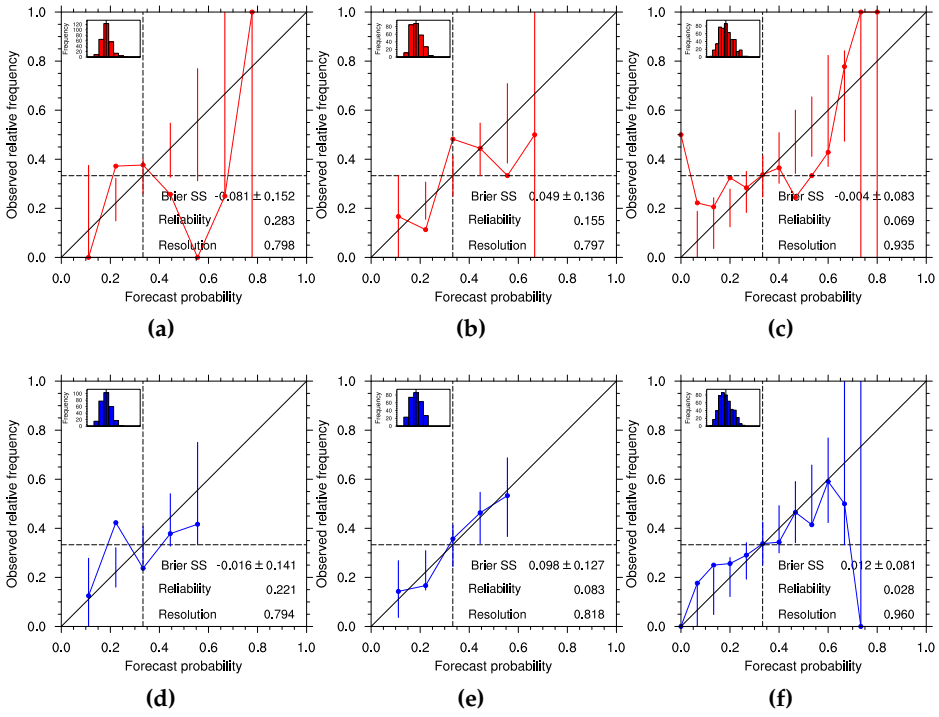


Figure B.8: Reliability of Kenya precipitation vs GPCP, details as in figure B.7.

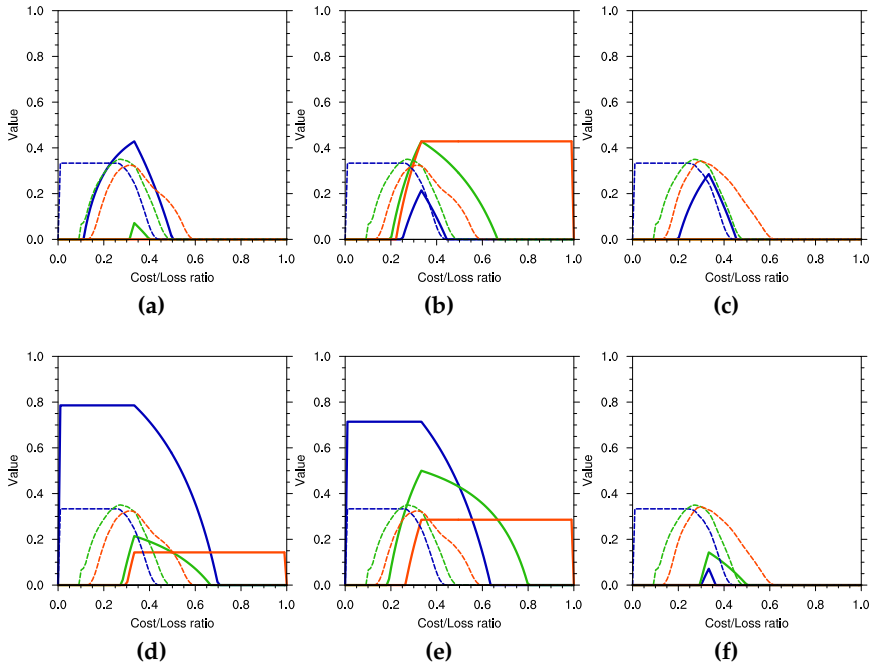


Figure B.9: Value of upper (a-c) and lower (d-f) tercile JAS temperature forecasts over Kenya, for DEMETER (a & d), ENSEMBLES (b & e) and System 4 (c & f). Curves for 30%, 50% and 70% decision thresholds are shown by blue, green and orange lines with dashed lines indicating 95% significance level for each threshold. Forecasts issued at the start of February.

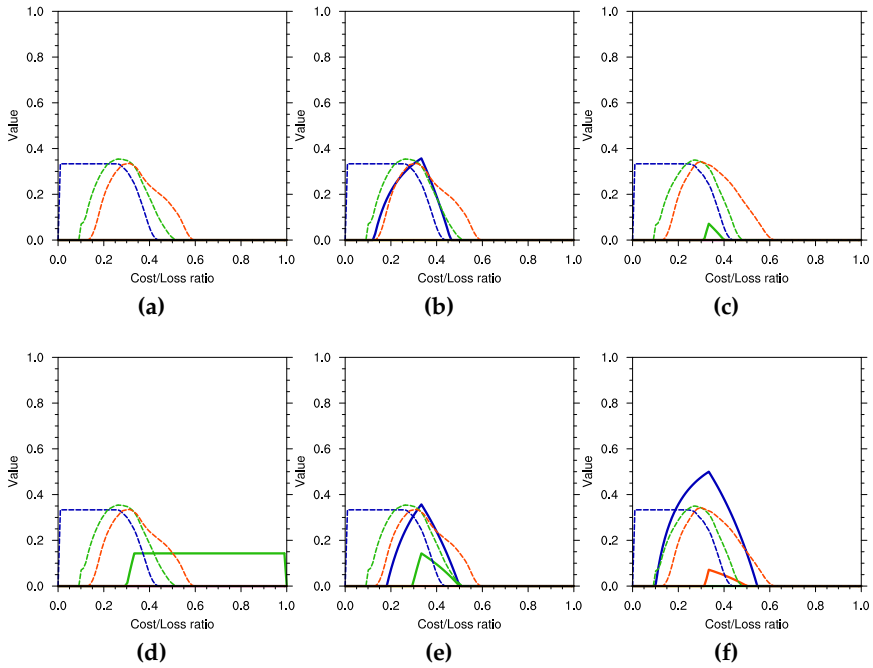


Figure B.10: Value of Bangladesh precipitation vs GPCP, details as in figure B.23.

in System 4. BSS for System 4 is also high for lower tercile forecasts, though it is lower than ENSEMBLES. This is because whilst the reliability of forecasts is high, System 4 is let down by its poor resolution, seen in the higher (i.e. worse) resolution component of the BSS.

For precipitation (figure B.8), BSS can not be separated from zero for any system, for upper or for lower tercile forecasts. For lower tercile forecasts the reliability component of ENSEMBLES and System 4 is relatively good, but both systems have a poor resolution component.

Finally value plots for Kenya temperature and precipitation are shown in figures B.9 and B.10. For lower tercile temperature DEMETER and ENSEMBLES have significant value at the 30% threshold, whilst for upper tercile events, the magnitude is only just over significance. System 4 has no significant value for either upper or lower tercile events. For precipitation, the value of is generally below significance for upper and lower tercile, except for System 4 which has value just above significance for lower tercile events.

B.2 Indian Subcontinent

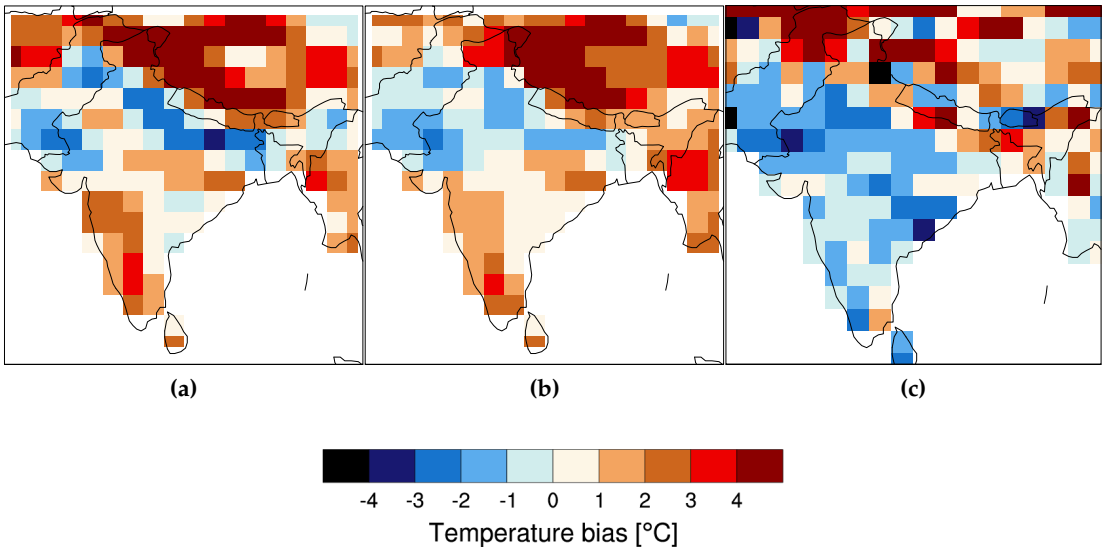


Figure B.11: Ensemble mean temperature bias, JJA over the Indian Subcontinent vs NCEP, for DEMETER, ENSEMBLES and System 4 (a-c).

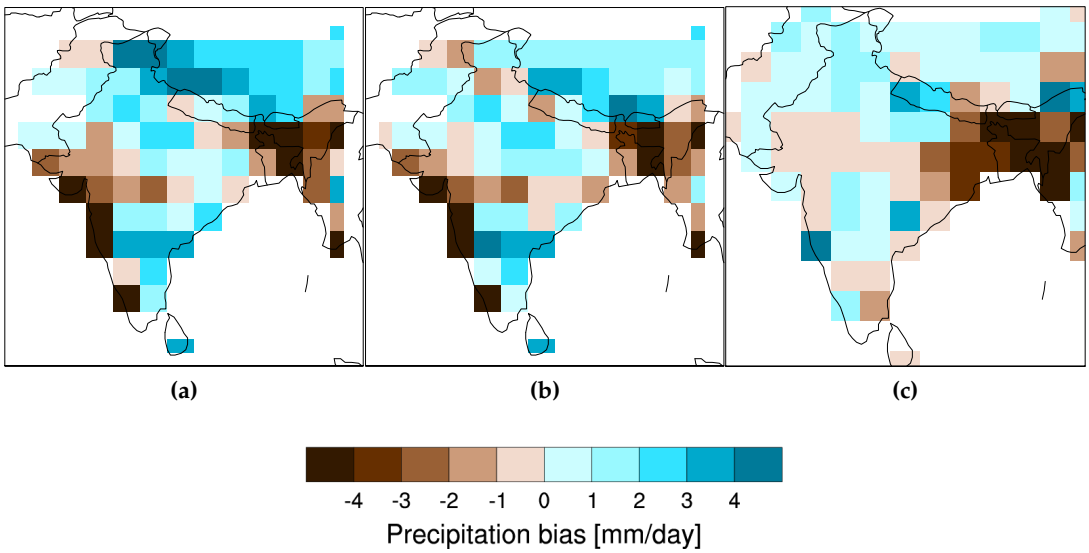


Figure B.12: As figure B.11, for ensemble mean precipitation vs GPCP. Only points are shown where the observed seasonal climatology is greater than 1mm/day.

Turning now to climate predictions for the start of the monsoon over the Indian subcontinent (JJA), figure B.11 shows the temperature bias over this region. The pattern for DEMETER and ENSEMBLES is quite similar, with a warm bias at the tip of India, a cold bias south of the Himalayas and a strong warm bias over the Tibetan plateau. The magnitude of the bias is slightly decreased in ENSEMBLES, but not by much. For

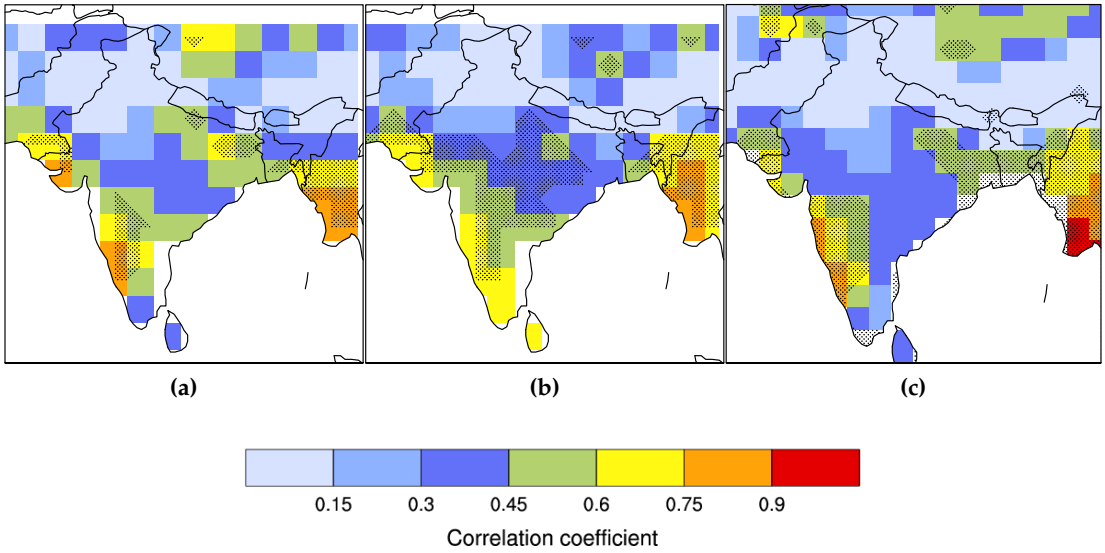


Figure B.13: Pearson's product-moment correlations of JJA ensemble mean precipitation over the Indian subcontinent vs NCEP, for DEMETER, ENSEMBLES and System 4 (a-c). Forecasts issued at the start of May. Stippled area shows 99% confidence level, with sample size adjusted to take account for autocorrelation.

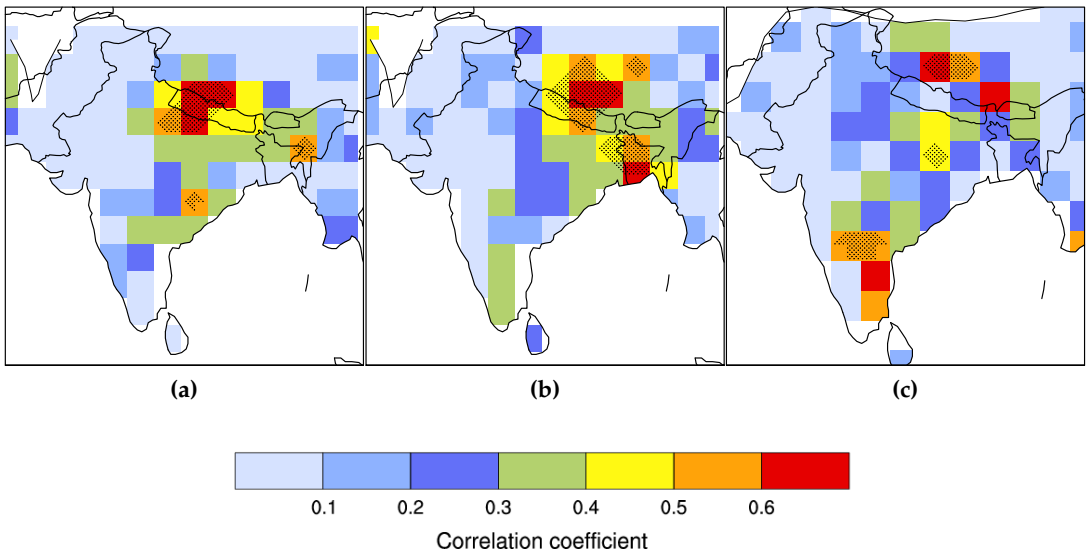


Figure B.14: As figure B.13, for ensemble mean precipitation vs GPCP. Only points are shown where the observed seasonal climatology is greater than 1mm/day.

System 4 (figure B.11c), the pattern is totally different; there is a cold bias of around 1 - 2 degrees C over most of India.

The similarity of DEMETER and ENSEMBLES over India is also apparent in the precipitation bias (figure B.12a and B.12b); there is a dry bias on the east side of India, a

wet bias in the South East, strong wet bias over the Himalayas and a dry bias over Bangladesh. There is only slight improvement between DEMETER and ENSEMBLES. System 4 on the other hand has a much reduced bias for most of India (figure B.12c), with the bias over most of the country not exceeding 1mm/day. The dry bias over Bangladesh however is present also in System 4 at the same magnitude.

Correlations for temperature and precipitation are shown in figures B.13 and B.14. The pattern for all models for temperature is similar, with a patch of significance in the south west of India, and another patch in the East, over Myanmar. The area of significance does increase between DEMETER and ENSEMBLES, spreading further inland from the west, whilst the area of significant temperature correlation for System 4 is more similar to DEMETER than ENSEMBLES.

Precipitation correlations show a similar pattern between the three systems, with DEMETER and ENSEMBLES showing most similarity. Only a few points have correlations above significance, with the highest correlations around Nepal and Bangladesh. System 4 is similar, except for a region in the east of the tip of India that has an increased correlation coefficient (figure B.14c).

ROC AUC for temperature and precipitation are shown in figures B.15 and B.16. The pattern of high ROC AUC generally follows the pattern of correlation, as it does for West Africa. For temperature, the highest significant ROC AUC for all systems is on the west coast of India, and around Myanmar. There is a slight increase between DEMETER and ENSEMBLES, with upper tercile forecasts generally better than lower tercile. The ROC AUC for System 4 does not indicate that it has any improvement on ENSEMBLES, but the regions of significance are similar, with a large coherent patch in the west coast of India, and some high score over Myanmar.

For precipitation (figure B.16), the scores are around 0.5 for most of the region, indicating that in these places the forecasting systems perform no better than climatology. The one patch of significant ROC AUC is over Nepal for DEMETER and ENSEMBLES, particularly for lower tercile forecasts. For System 4 this score is reduced, though it does have a small patch of significant skill over the east of the tip of India, which is slightly higher for upper tercile forecasts.

Turning now to the smaller regions and reliability of temperature forecasts for the west India region (figure B.17). System 4 shows a positive BSS for upper and lower tercile temperature forecasts. These is also reliability, as all the points lie inside the consistency bars. The reliability component of the BSS is also small. In comparison DEMETER and ENSEMBLES have poor reliability components and the points lie outside the consistency bars. Despite this, forecasts for upper tercile temperature have the lowest

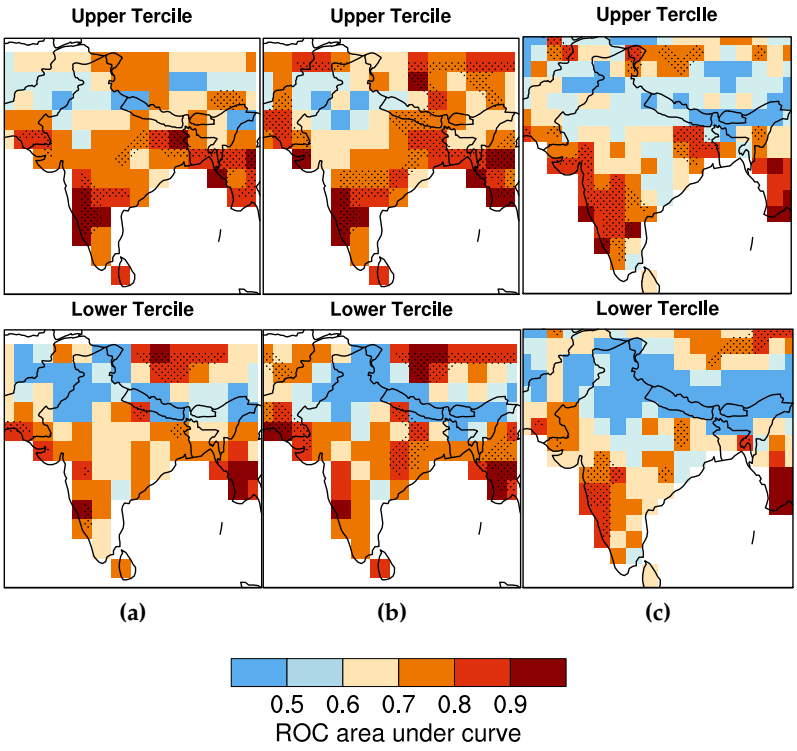


Figure B.15: Relative operating characteristic area under curve (ROC AUC) for JJA temperature vs NCEP, for the May start dates of DEMETER, ENSEMBLES and System 4 (a-c). Stippled area indicates where the AUC is significant at the 95% level.

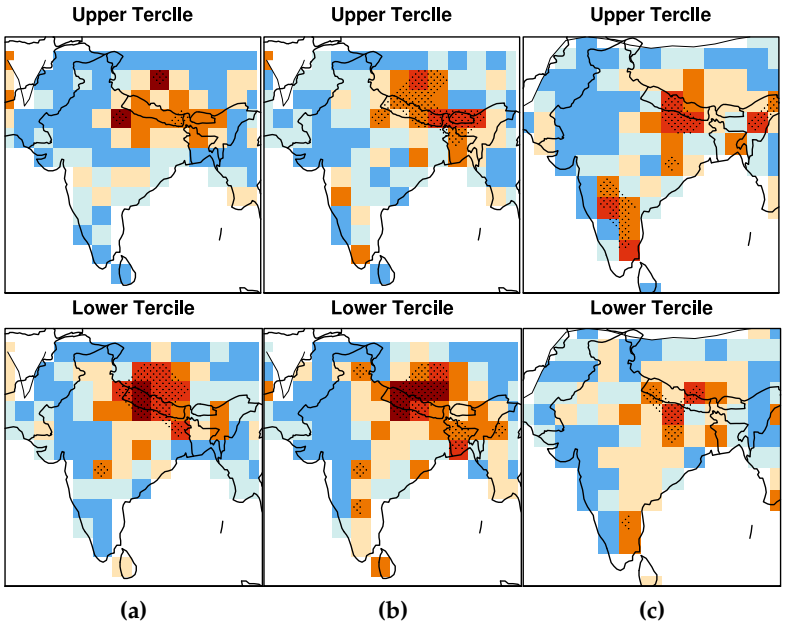


Figure B.16: As figure B.15, for precipitation vs GPCP. Only points are shown where the observed seasonal climatology is greater than 1mm/day.

BSS for System 4. This is due to the resolution of both DEMETER and ENSEMBLES being better than System 4 and correspondingly improving their BSS. For lower tercile events all forecast systems have a positive BSS, though there is a large error for DEMETER and ENSEMBLES, suggesting particularly for DEMETER that it cannot be distinguished meaningfully from a climatological score of zero. The sharpness of System 4 forecasts is also improved over DEMETER and ENSEMBLES, with more high probability forecasts.

For precipitation (figure B.18), the BSS for all events and systems cannot be distinguished from zero. This, along with the poor reliability diagrams, suggests that the precipitation forecasts here are unreliable and have not improved significantly since DEMETER.

Temperature forecasts over west India have a high value above significance for all systems, as demonstrated in figure B.19, though this has remained roughly constant between DEMETER and System 4. For precipitation (figure B.19) there is a slight positive value for upper tercile forecasts, this is only slightly greater than zero. For lower tercile forecasts the value is lower and only slightly above significance, and has also remained roughly constant between the systems.

For precipitation (figure B.20) only ENSEMBLES has any value above significance, for upper tercile forecasts using the 50% decision threshold (figure B.20b). The magnitude of the value is not high however. For lower tercile no system shows above significance and for most thresholds the value is below zero.

Turning now to Bangladesh and reliability of temperature forecasts (figure B.21). Reliability curves sit generally within the consistency bars for upper tercile forecasts for all systems, however the BSS for upper tercile prediction are not significantly different from zero, except for DEMETER upper tercile. System 4 is generally reliable for forecast probabilities less than 0.8, but above this it performs poorly. For lower tercile temperature prediction, the BSS is below climatology for all systems.

Looking at the reliability diagrams for precipitation forecasts over Bangladesh (figure B.22), the reliability curves generally lie within the consistency bars, as would be expected from a reliable system, though the consistency bars are large. Also the BSS is not significantly above zero for any of the systems for upper nor lower precipitation, suggesting that precipitation forecasts are poor here.

Finally value curves for temperature and precipitation forecasts over Bangladesh are shown in figures B.23 and B.24. Value for all systems is only just above significance for upper and lower tercile temperature forecasts. Precipitation value is also not large though ENSEMBLES and System 4 have some value above significance for lower tercile events, at the 30 and 50% thresholds respectively.

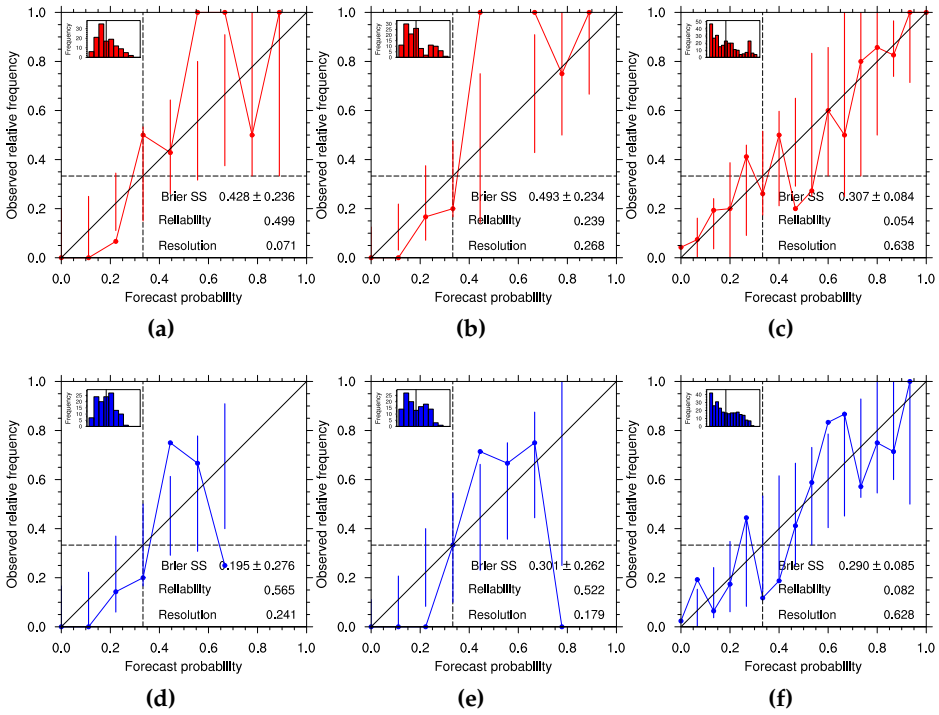


Figure B.17: Reliability of upper (a-c) and lower (d-f) tercile JJA temperature forecasts over west India (vs NCEP): DEMETER (a & d), ENSEMBLES (b & e) and System 4 (c & f). Forecasts issued at the start of May.

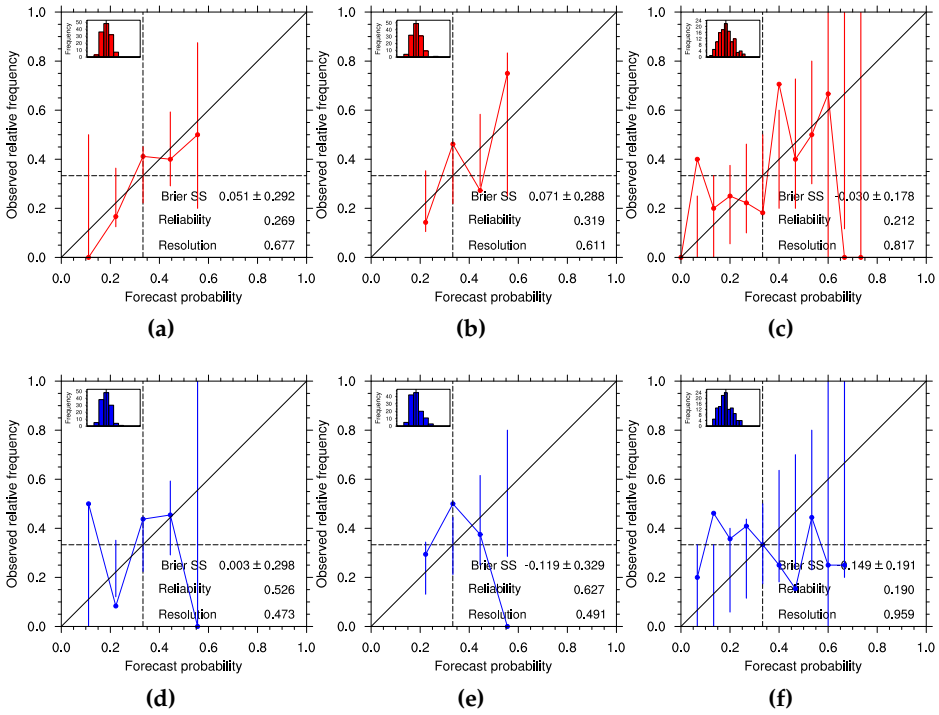


Figure B.18: Reliability of west India precipitation vs GPCP, details as in figure B.17.

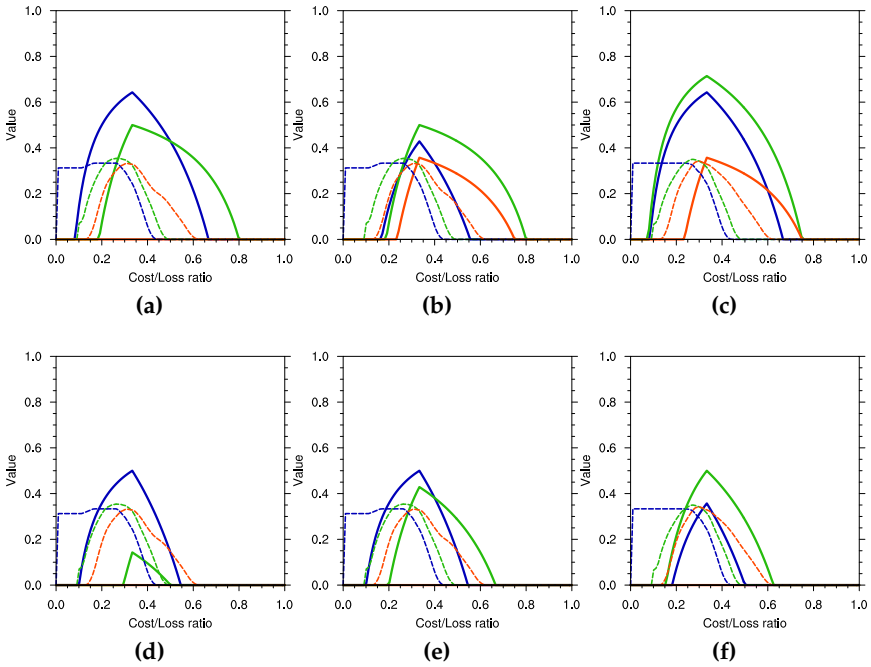


Figure B.19: Value of upper (a-c) and lower (d-f) tercile JAS temperature forecasts over west India, for DEMETER (a & d), ENSEMBLES (b & e) and System 4 (c & f). Curves for 30%, 50% and 70% decision thresholds are shown by blue, green and orange lines with dashed lines indicating 95% significance level for each threshold. Forecasts issued at the start of May.

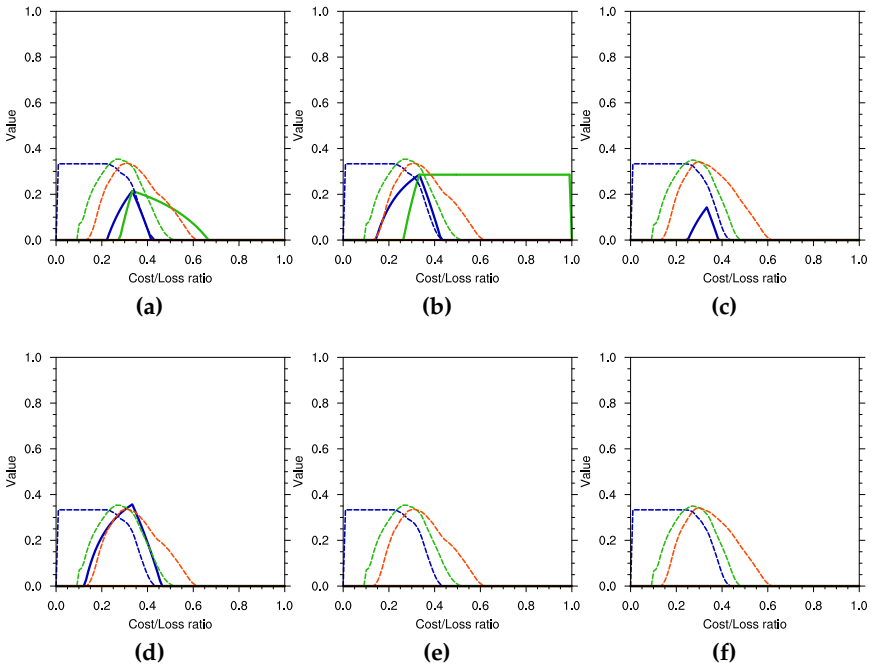


Figure B.20: Value of west India precipitation vs GPCP, details as in figure 5.24.

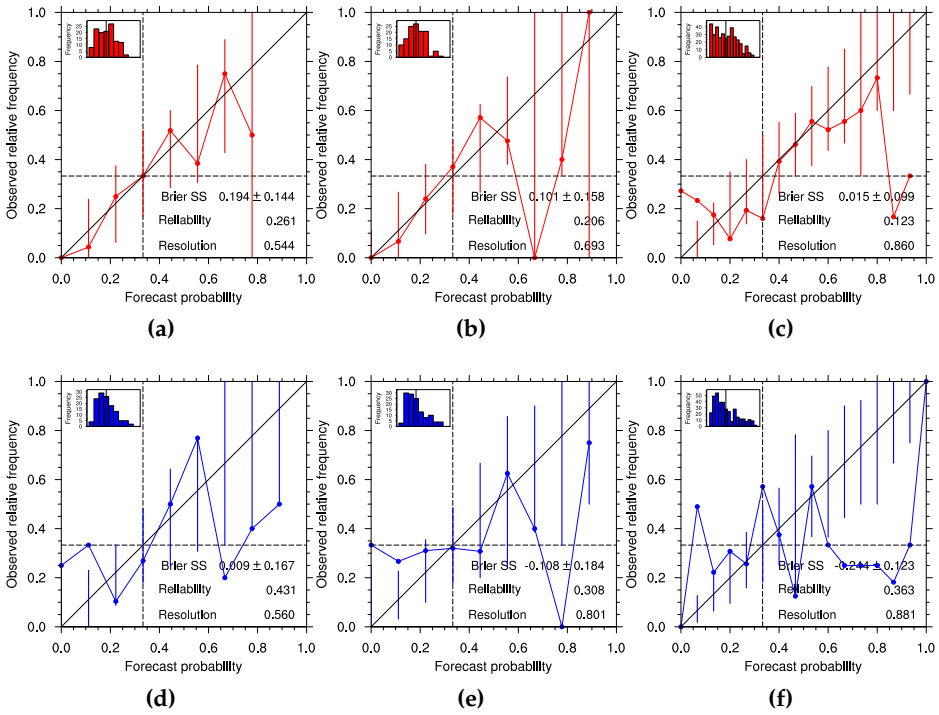


Figure B.21: Reliability of upper (a-c) and lower (d-f) tercile JJA temperature forecasts over Bangladesh (vs NCEP): DEMETER (a & d), ENSEMBLES (b & e) and System 4 (c & f). Forecasts issued at the start of May.

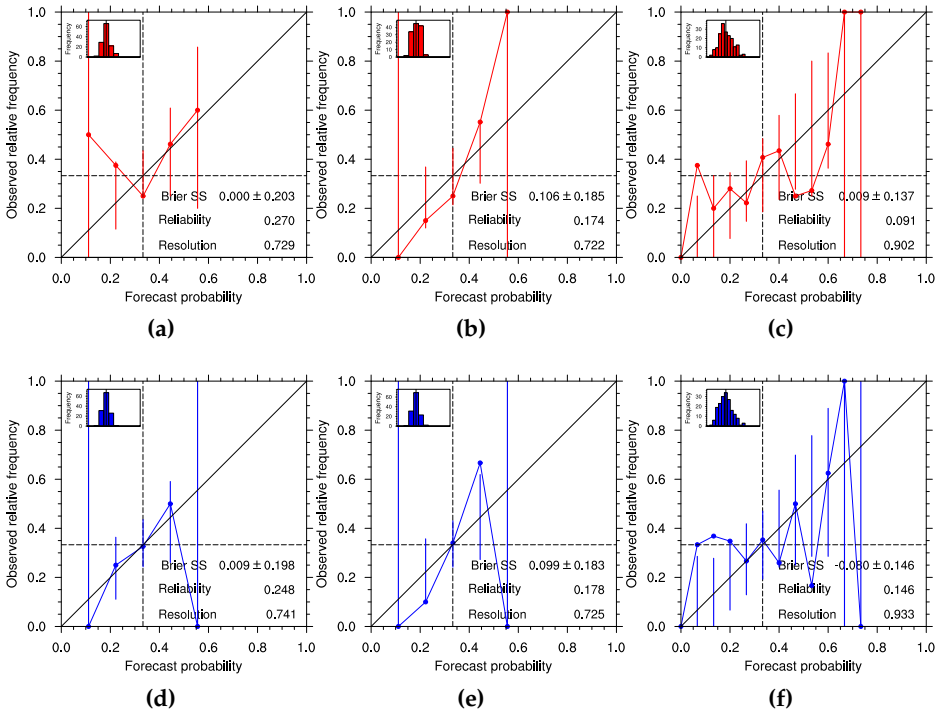


Figure B.22: Reliability of Bangladesh precipitation vs GPCP, details as in figure B.21.

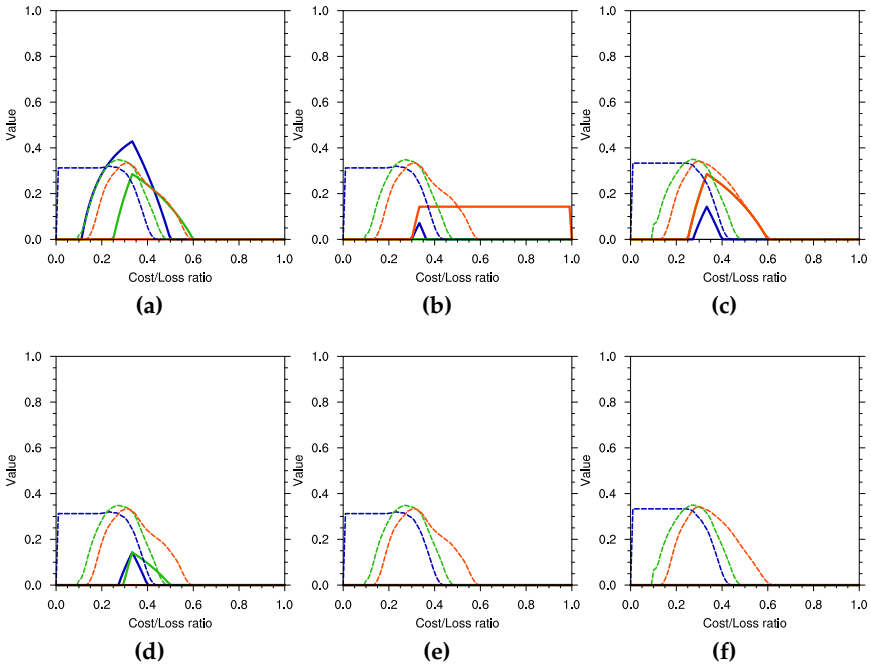


Figure B.23: Value of upper (a-c) and lower (d-f) tercile JAS temperature forecasts over Bangladesh, for DEMETER (a & d), ENSEMBLES (b & e) and System 4 (c & f). Curves for 30%, 50% and 70% decision thresholds are shown by blue, green and orange lines with dashed lines indicating 95% significance level for each threshold. Forecasts issued at the start of May.-

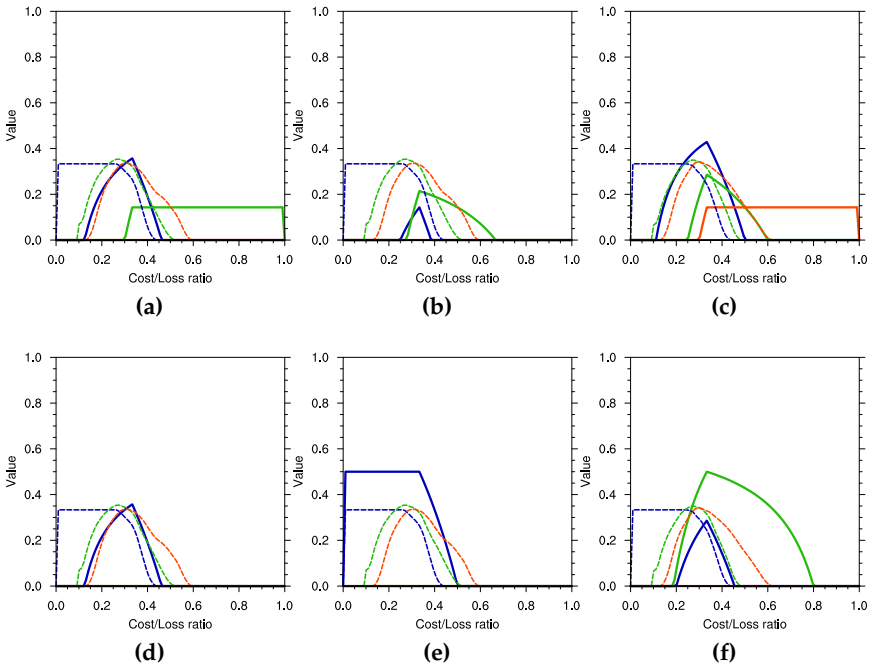


Figure B.24: Value of Bangladesh precipitation vs GPCP, details as in figure B.23.

APPENDIX C

Extra figures for Chapter 6

Contained in this chapter are extra figures for chapter 6, for East Africa and for the Indian subcontinent.

C.1 East Africa

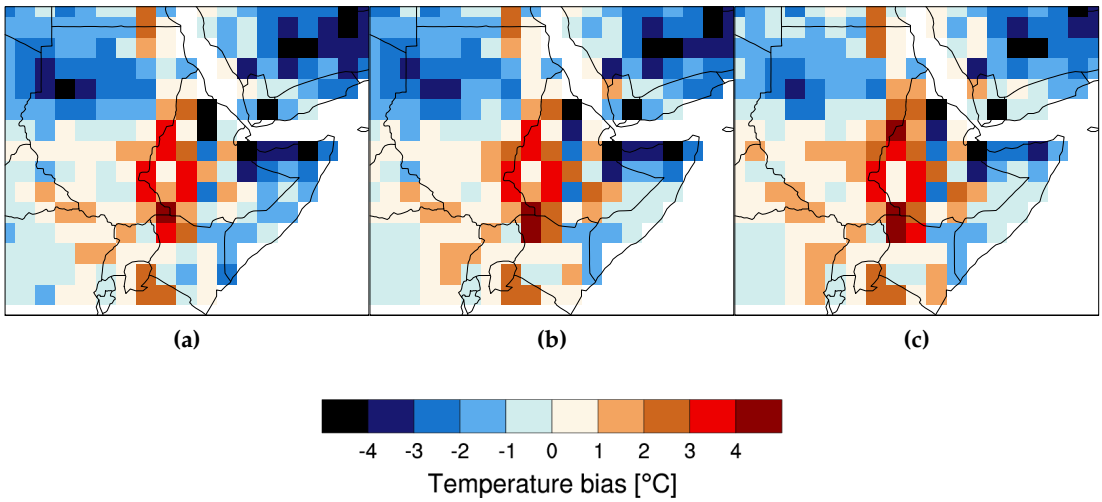


Figure C.1: Ensemble mean MAM average temperature bias over the East Africa vs NCEP, for System 4 forecasts issued November, January and March (a-c).

Biases for average temperature over East Africa during MAM are shown in figure C.1. The bias does not change significantly between start dates. There is a warm bias of over three degrees in western Ethiopia, with a cold bias over the Sahara and over Somalia. Elsewhere the bias is under one degree. For precipitation (figure C.1 the dry bias is low, under one degree for most of the region. There is little difference in the bias between the start dates.

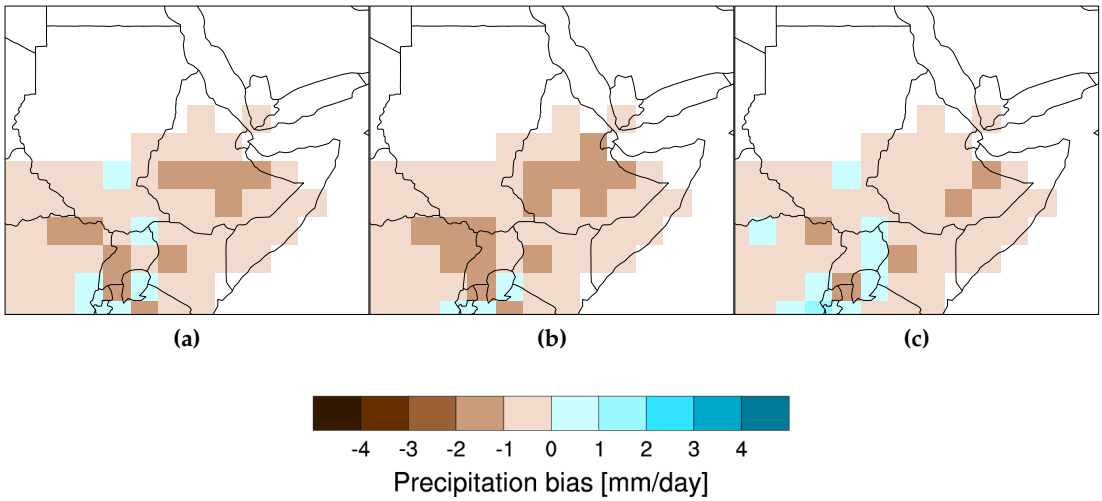


Figure C.2: As figure C.1, for ensemble mean precipitation vs GPCP. Only points are shown where the observed seasonal climatology is greater than 1mm/day.

Correlations for temperature are shown in figure C.3. There is significant correlation around the coast, with a maximum in the north east Congo and over Somalia. There is no improvement as the target is approached; in fact the value of the correlation coefficient is lower for March forecasts than it is for those initialised in November. For precipitation (figure C.4 the correlation does improve with lead time; November forecasts have no significant correlation anywhere, whilst the value for March forecasts is higher. Significant correlations are observed over the south east of the region, though only for a few grid points.

Maps of ROC AUC for temperature are shown in figure C.5. Longest lead forecasts have significant skill everywhere except for the desert, whilst this skill slightly reduces by March. For precipitation (figure C.6) the longest lead time forecasts generally have low ROC AUC, under 0.6, whilst the March forecasts have a patch of significant score over Kenya, particularly for upper tercile events.

Reliability plots for Kenya are shown in figure C.7. The BSS is positive and highest at the longest lead time and it reduces as the target is approached, whilst the reliability for both upper and lower tercile forecasts is highest in November. For precipitation (figure C.8) BSS is zero for lead times longer than a month, whilst forecasts issued at the start of the rainy season have a reliability curve laying inside the consistency bars, as well as a positive BSS.

Value curves for Kenya temperature forecasts is shown in figure C.9. The value of lower tercile forecasts is positive and similar for all lead times, whilst the value of upper tercile forecasts is greatest in November and February. Forecasts for upper tercile temperature

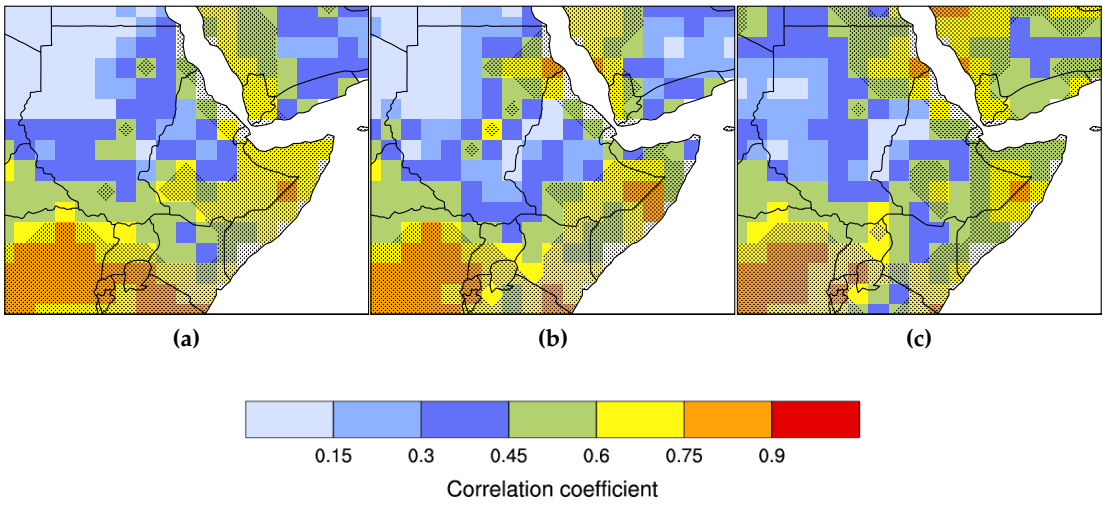


Figure C.3: Pearson's product-moment correlations of MAM ensemble mean temperature vs NCEP, for System 4 forecasts issued November, January and March (a-c). Stippled area shows 99% confidence level, with sample size adjusted to take account for autocorrelation.

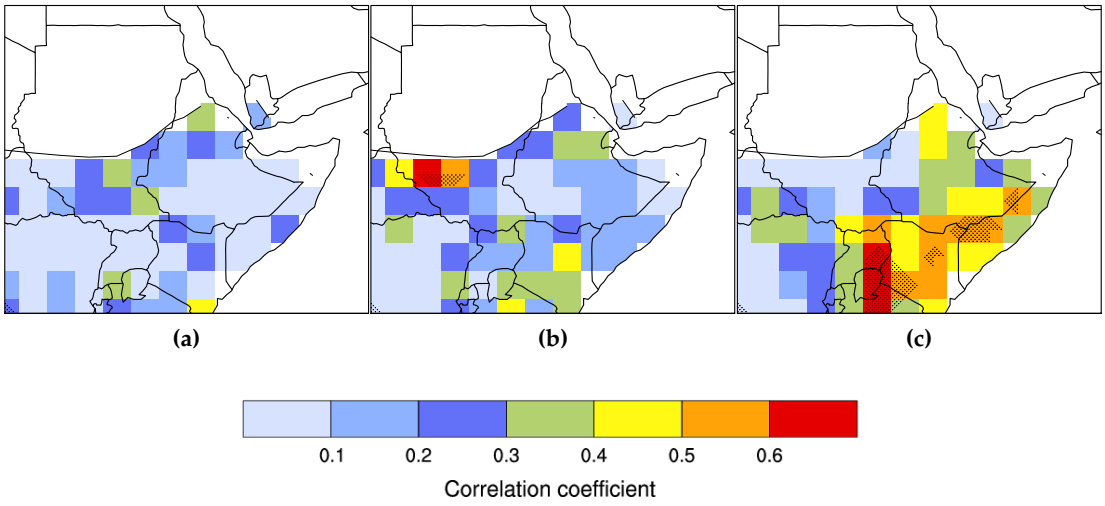


Figure C.4: As figure C.3, for ensemble mean precipitation vs GPCP. Only points are shown where the observed seasonal climatology is greater than 1mm/day.

initialised at the start of the rainy season have the lowest value.

For precipitation (figure C.10), there is no value above significance for upper tercile forecasts, except for forecasts issued at the start of the rainy season, though this is only just above significance. For lower tercile forecasts there is no value in forecasts initialised in November or in February, whilst there is a high value at the 30% decision threshold for March forecasts, with a smaller value for the 50 and 70% thresholds.

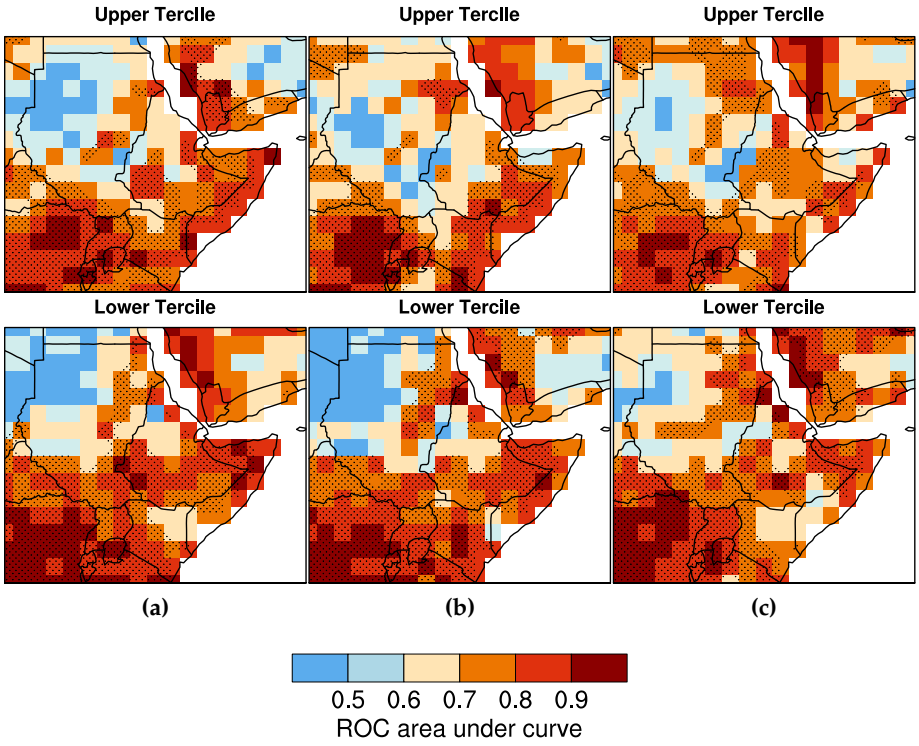


Figure C.5: Relative operating characteristic area under curve (ROC AUC) for MAM temperature vs NCEP, for System 4 forecasts issued November, January and March (a-c). Stippled area indicates where the AUC is significant at the 95% level.

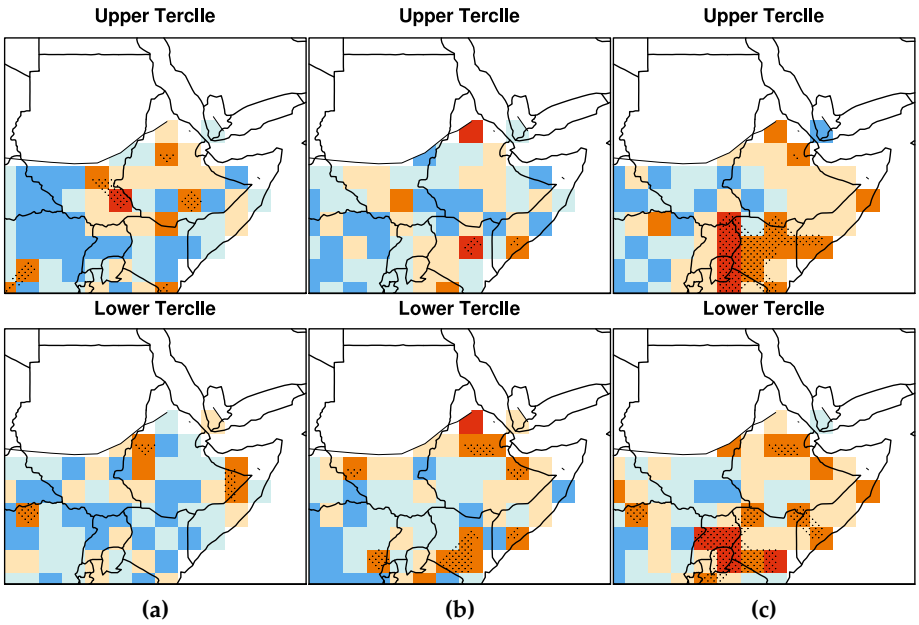


Figure C.6: As figure C.5, for precipitation vs GPCP. Only points are shown where the observed seasonal climatology is greater than 1mm/day.

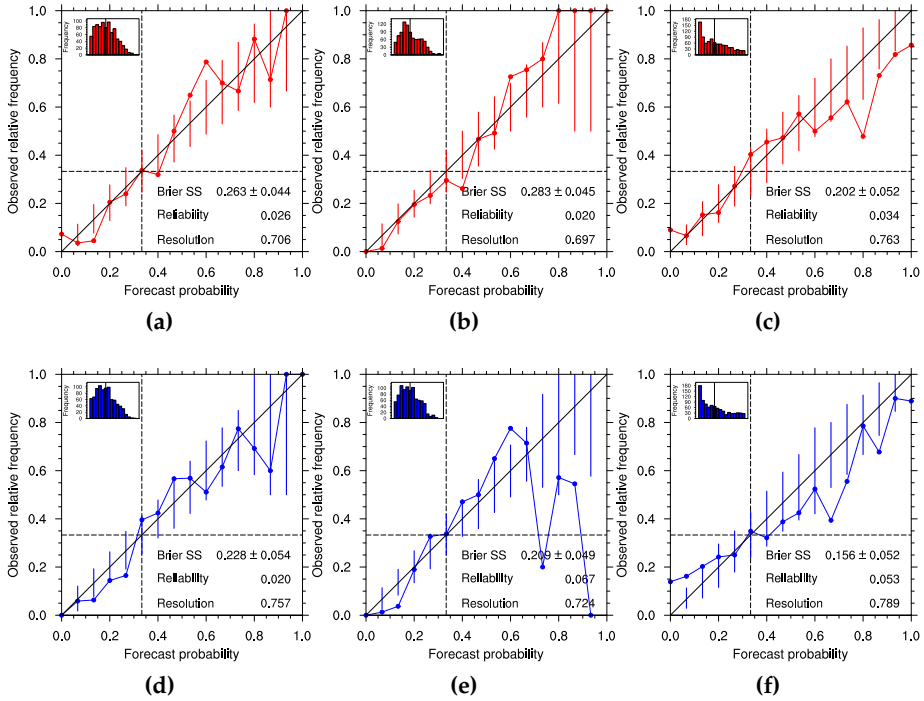


Figure C.7: Reliability of upper (a-c) and lower (d-f) tercile MAM temperature forecasts over Kenya. Reliability is shown for System 4 forecasts issued November (a & d), January (b & e) and March (c & f).

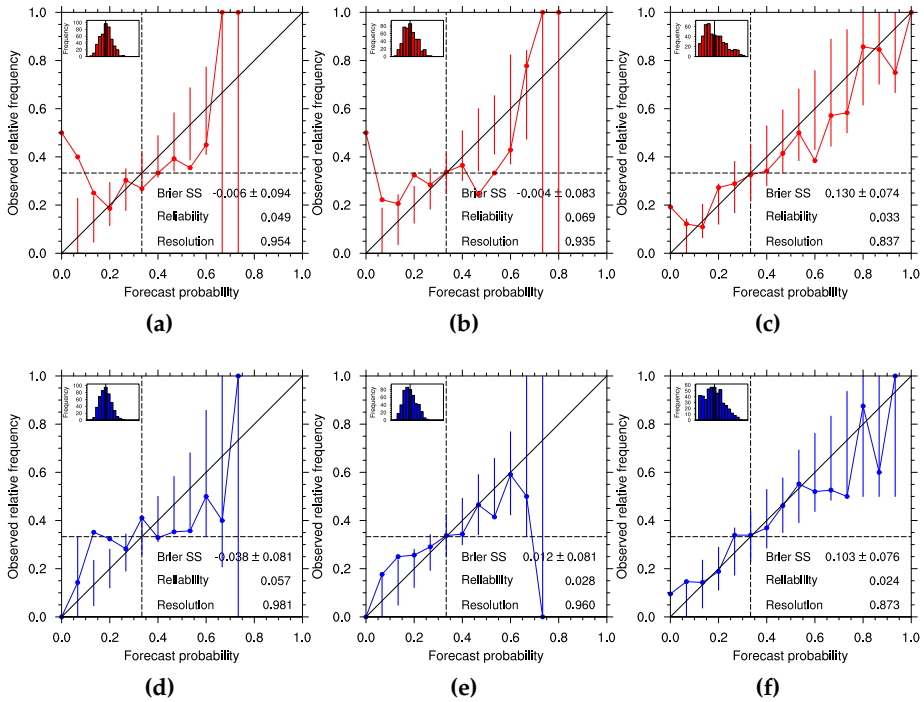


Figure C.8: Reliability of Kenya precipitation vs GPCP, details as in figure C.7.

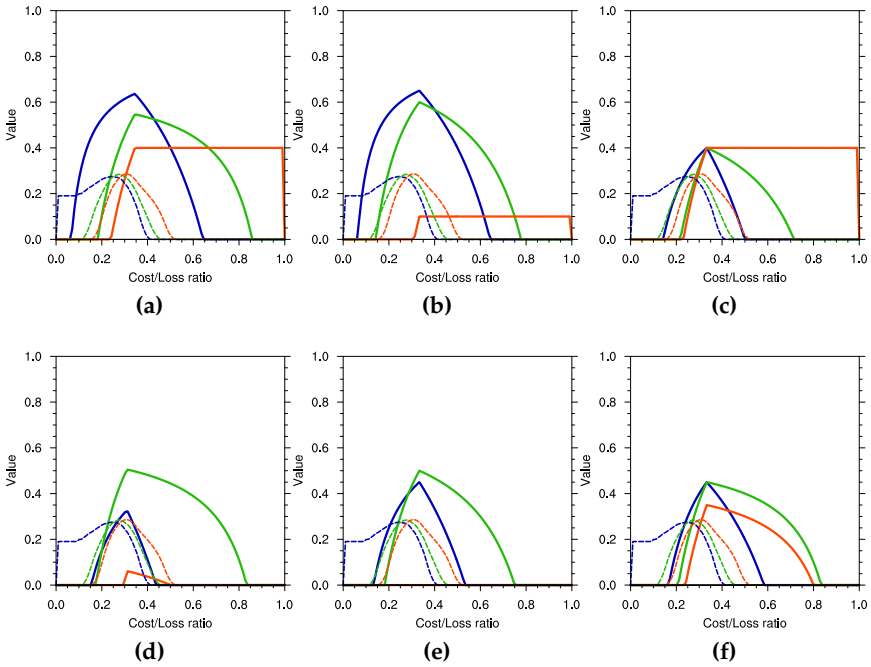


Figure C.9: Value of upper (a-c) and lower (d-f) tercile JAS temperature forecasts over Kena, for System 4 forecasts issued November (a & d), January (b & e) and March (c & f). Curves for 30%, 50% and 70% decision thresholds are represented by blue, green and orange lines with dashed lines indicating 95% significance level for each threshold.

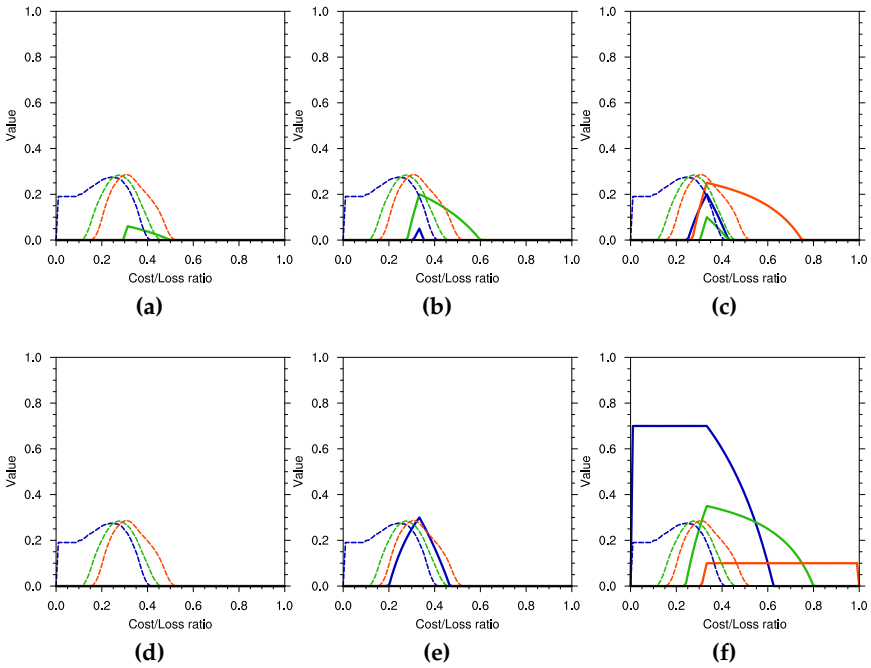


Figure C.10: Value of Bangladesh precipitation vs GPCP, details as in figure C.9.

C.2 Indian Subcontinent

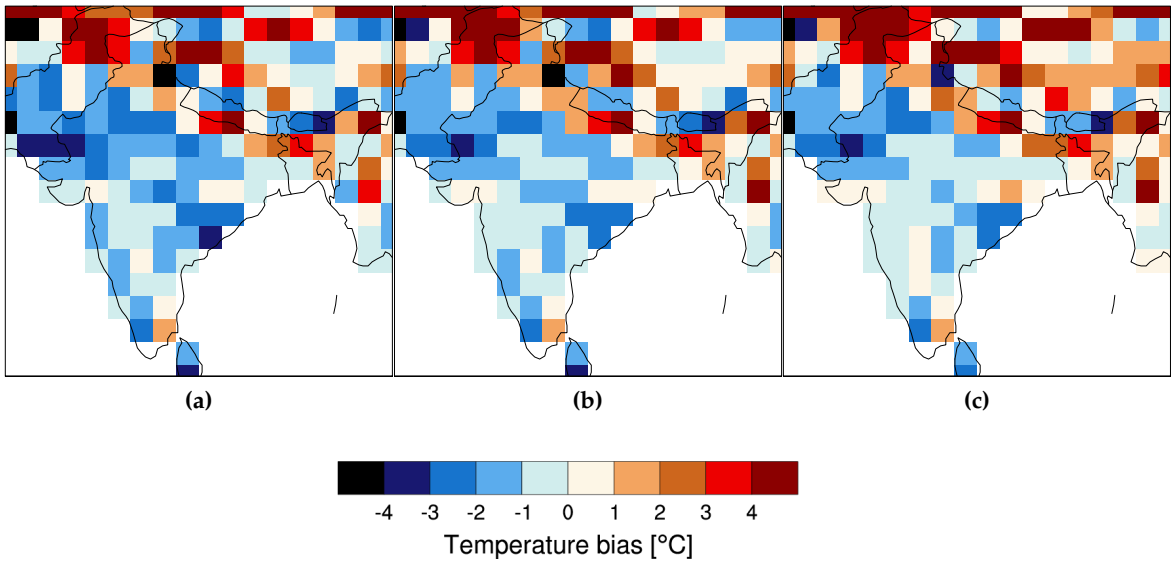


Figure C.11: Ensemble mean JJA average temperature bias over the Indian subcontinent vs NCEP, for System 4 forecasts issued February, April and June (a-c).

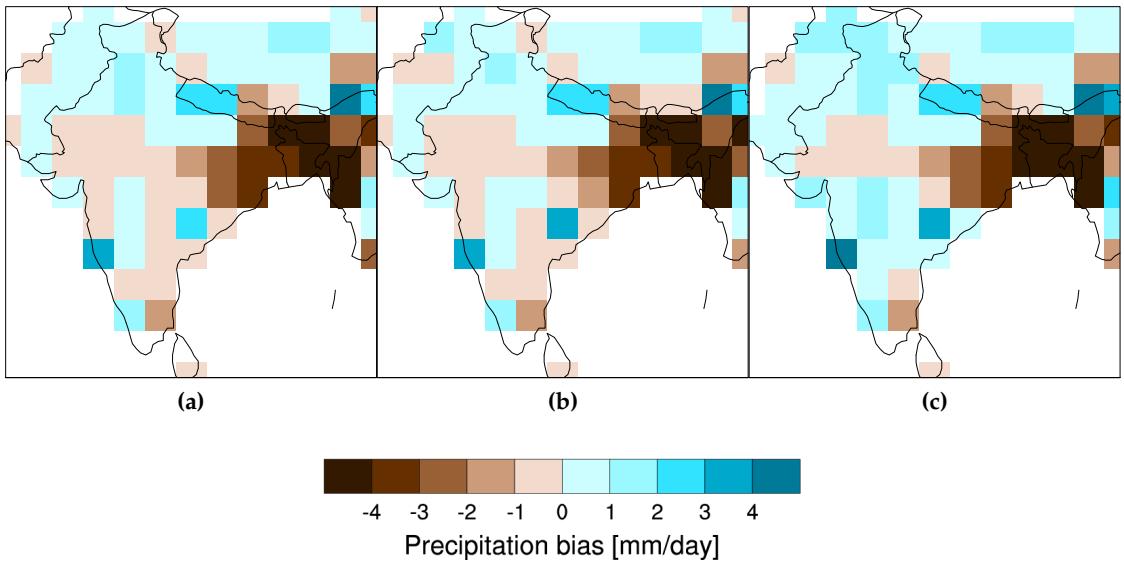


Figure C.12: As figure C.11, for ensemble mean precipitation vs GPCP. Only points are shown where the observed seasonal climatology is greater than 1mm/day.

The lead time dependent temperature bias for the Indian subcontinent is shown in figure C.11. The bias is generally cold for the whole of India, with a maximum around the north west, near to the coast of Pakistan. Around the Himalayas and further north the bias is complex and not generally coherent, suggesting a problem simulating the climate in a

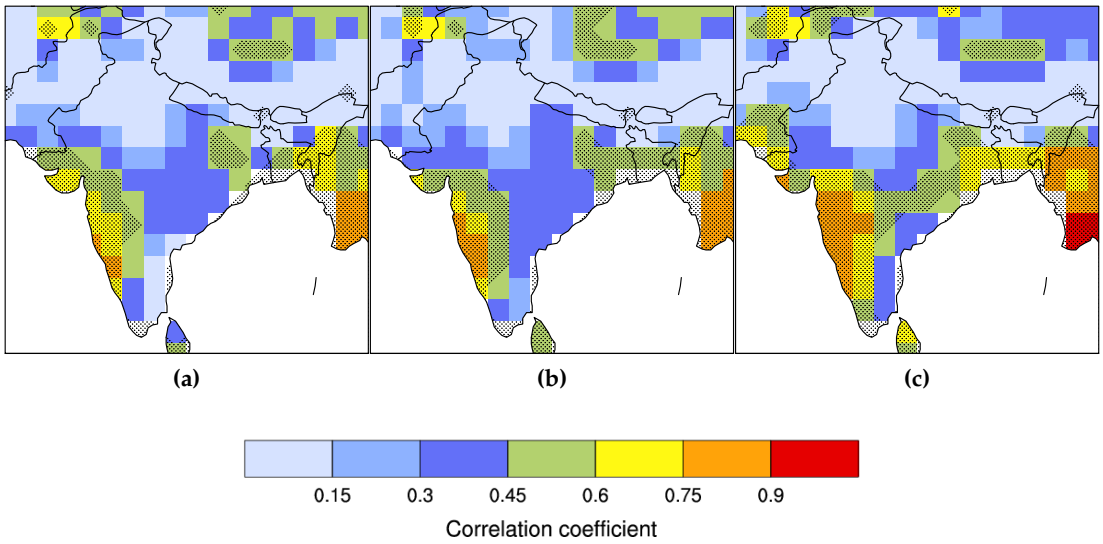


Figure C.13: Pearson's product-moment correlations of JJA ensemble mean temperature vs NCEP, for System 4 forecasts issued February, April and June (a-c). Stippled area shows 99% confidence level, with sample size adjusted to take account for autocorrelation.

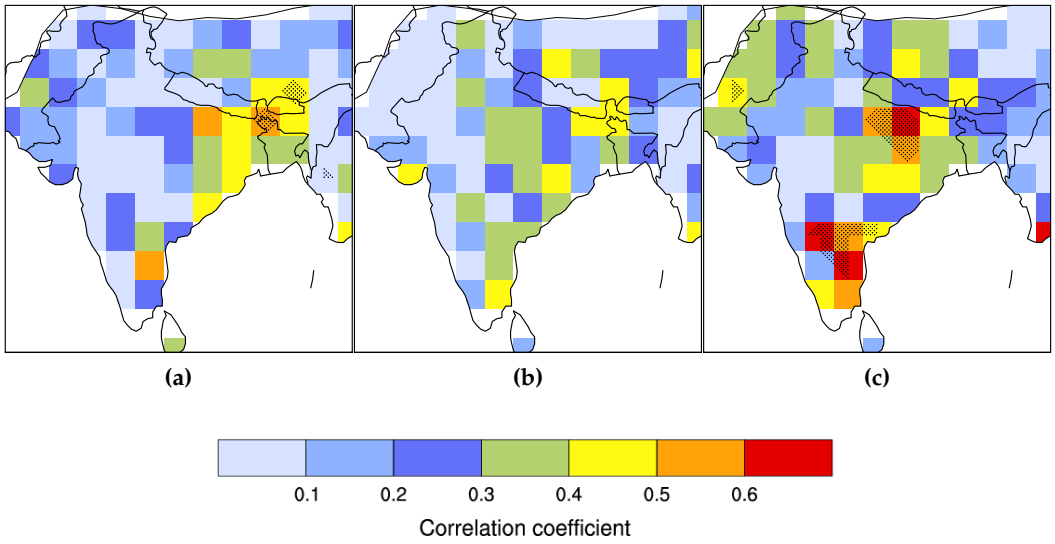


Figure C.14: As figure C.13, for ensemble mean precipitation vs GPCP. Only points are shown where the observed seasonal climatology is greater than 1mm/day.

region of complex topography (unsurprising considering the sharp gradients compared to the low resolution of the model). As the target is approached, the bias is slightly reduced, but not significantly.

For precipitation the bias is shown in figure C.12. The area of the highest bias is a region over Bangladesh, with the amount of rainfall too low by over four mm/day. Outside this

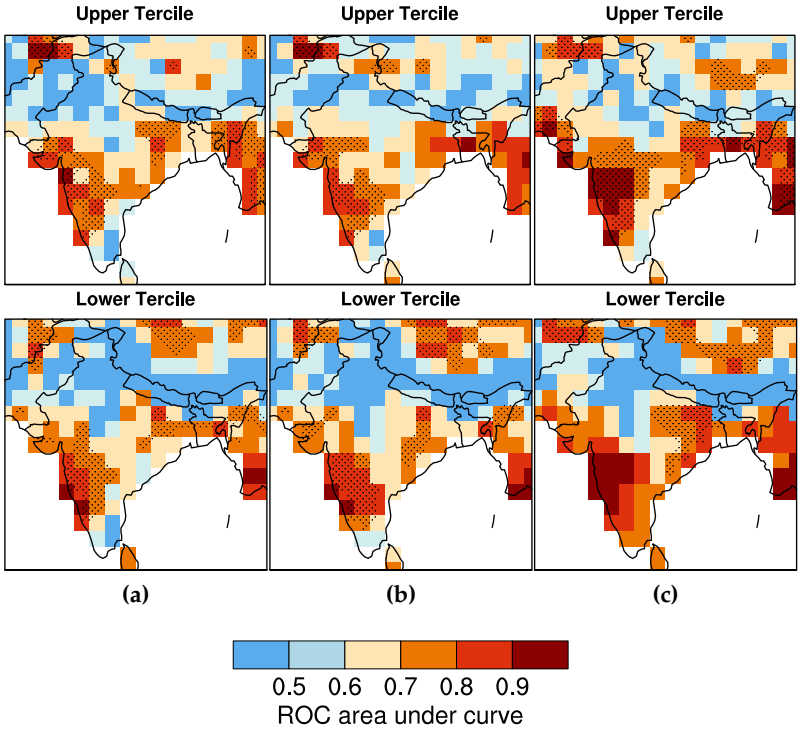


Figure C.15: Relative operating characteristic area under curve (ROC AUC) for JJA temperature vs NCEP, for System 4 forecasts issued February, April and June (a-c). Stippled area indicates where the AUC is significant at the 95% level.

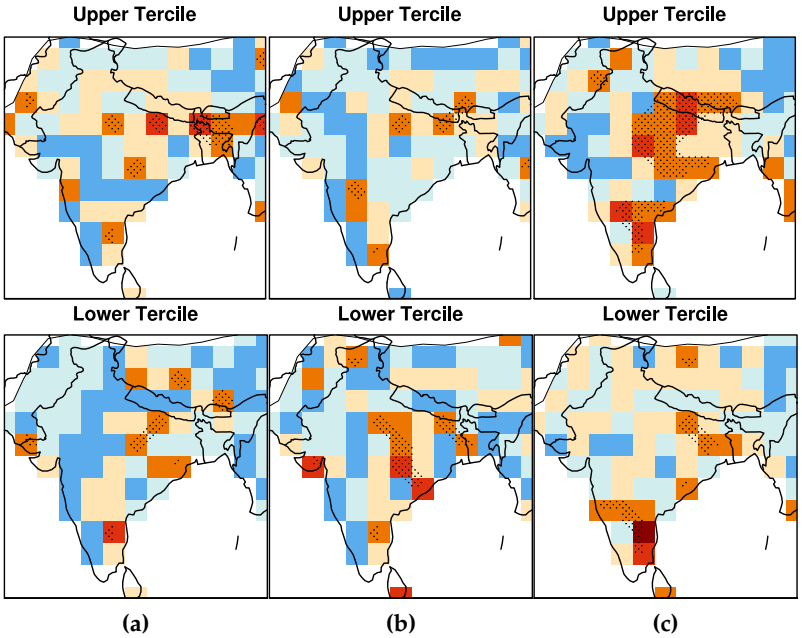


Figure C.16: As figure C.15, for precipitation vs GPCP. Only points are shown where the observed seasonal climatology is greater than 1mm/day.

region the bias is lower and generally under one mm/day. The precipitation bias with lead time is relatively static, with the only significant difference between the February and June forecasts being a reversal of the sign of the small bias over central south India.

For temperature correlations (figure C.13), there is no noticeable difference between forecasts made in February and April, with the area of significance limited to the west coast and to the Bangladesh/Myanmar area. Forecasts made in June have a higher value of the correlation coefficient in these regions, whilst the area of significant correlation is extended to join the two regions.

For precipitation (figure C.14), correlations are not significant anywhere for February forecasts. The values increase slightly for April whilst remaining below significance. In June the correlation is again higher and now significant, with a patch of skill in the east of the tip of India, and another further north, to the south of Nepal.

ROC AUC is shown for temperature in figure C.15. The pattern is similar to that for the correlation coefficient, with significant scores over the west coast of India, and over Myanmar. As the target is approached, ROC AUC increases steadily in these regions, with large areas where the score is over 0.9.

For precipitation (figure C.16) there are few grid points of significant skill for the February forecasts, for upper or lower tercile forecasts. This is similar for May forecasts, whilst in June the skill of forecasts is greater, particularly for upper tercile events; an area in the south east of India and another south of Nepal both have significant ROC AUC.

Reliability of temperature forecasts for West India are shown in figure C.17. BSS increases as the target is approached, with highest scores for June forecasts. The sharpness of the forecast distribution also increases significantly, with many forecasts in the 0 probability category.

For precipitation forecasts (figure C.18) the BSS is highest for June forecasts, though taking the error into account shows that it is not significantly different from zero. This is the case for upper and lower tercile events, and though the curve generally lies inside the consistency bars, the consistency bars are so wide (due to the small sample) that no strong conclusion can be drawn.

Value of temperature forecasts (figure C.19), is highest at the shortest lead time; June forecasts for upper and lower tercile events have a value high above significance. At longer lead times the value is less, with the February forecasts only just above significance.

For precipitation forecasts (figure C.20), value is below significance for all lead times,

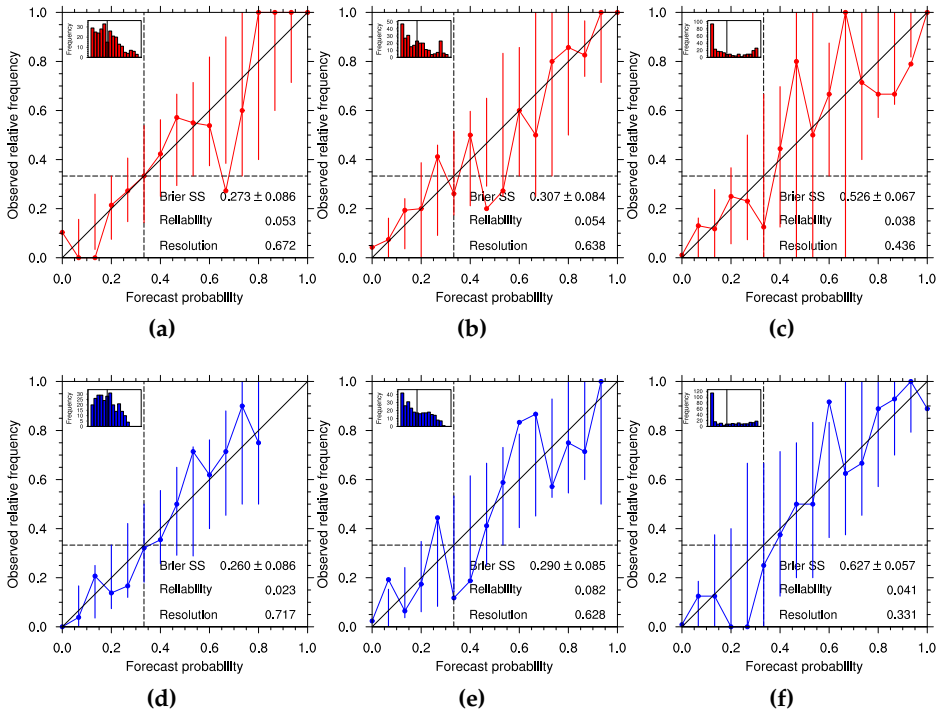


Figure C.17: Reliability of upper (a-c) and lower (d-f) tercile JJA temperature forecasts over the West India region. Reliability is shown for System 4 forecasts issued February (a & d), April (b & e) and June (c & f).

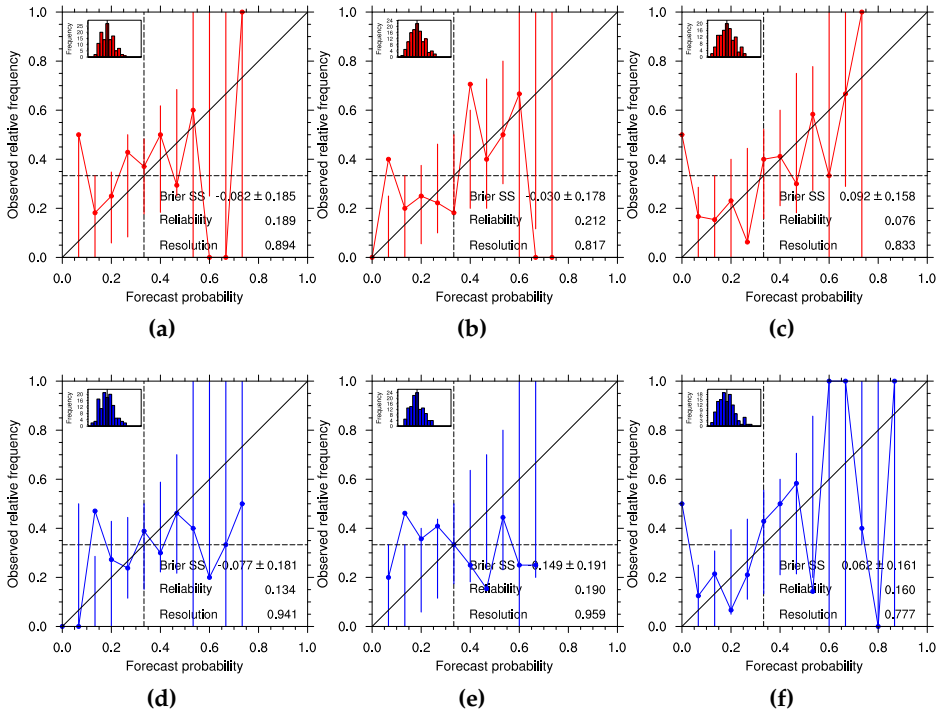


Figure C.18: Reliability of West India precipitation vs GPCP, details as in figure C.17.

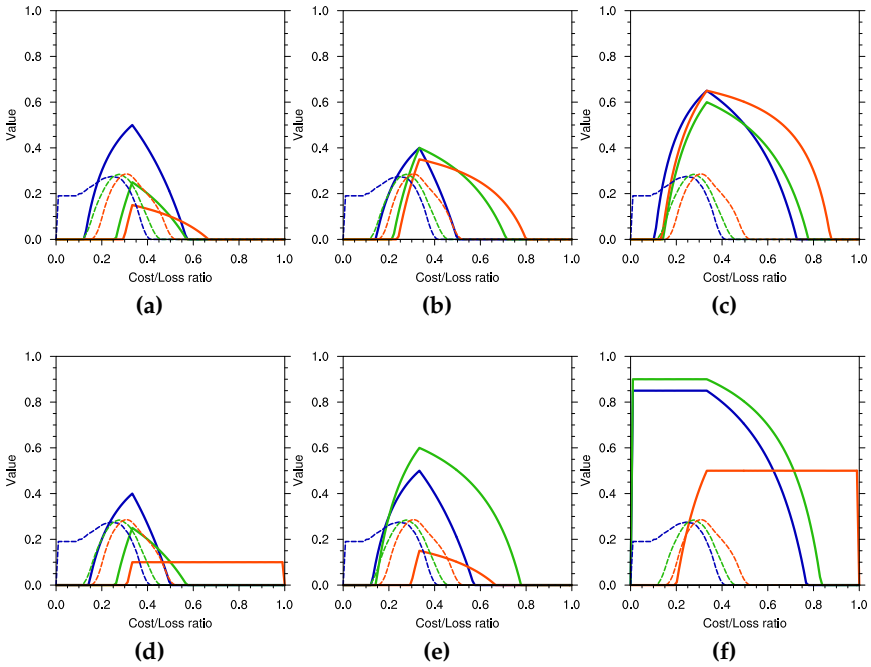


Figure C.19: Value of upper (a-c) and lower (d-f) tercile JAS temperature forecasts over West India, for System 4 forecasts issued February (a & d), April (b & e) and June (c & f). Curves for 30%, 50% and 70% decision thresholds are represented by blue, green and orange lines with dashed lines indicating 95% significance level for each threshold.

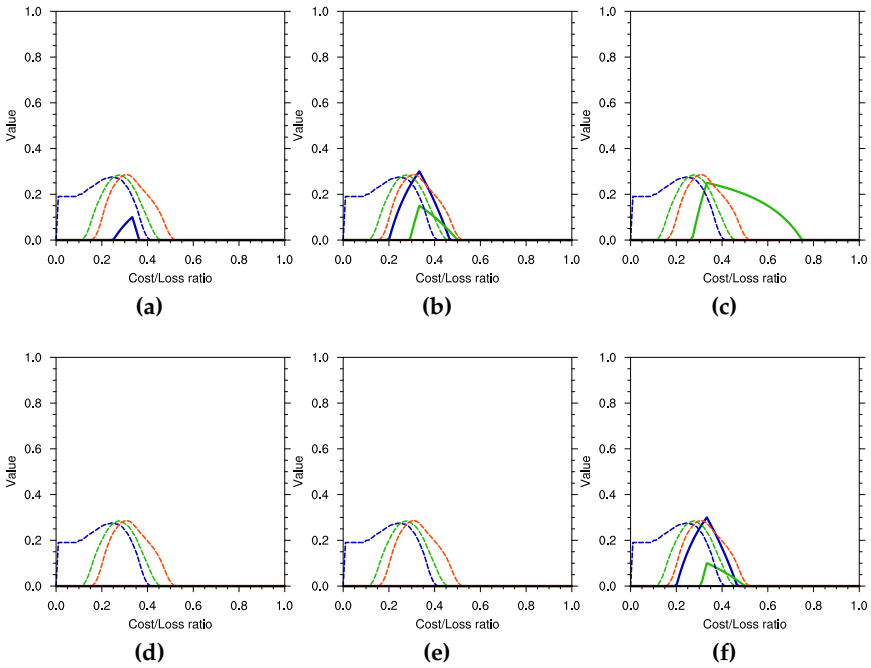


Figure C.20: Value of West India precipitation vs GPCP, details as in figure C.19.

with only one decision threshold curve for upper tercile June forecasts above the significance level. Forecasts of precipitation over West India then have limited value.

Temperature forecasts over Bangladesh have no reliability, as shown by the reliability curves in figure C.21, with zero or negative BSS for all lead times for upper and lower tercile events. For precipitation (figure C.22) the result is the same; BSS is zero or negative for all lead times for both event categories.

Temperature forecasts over Bangladesh have no value above significance for any lead time (figure C.23), whilst for precipitation (figure C.24) the value lays around the significance level, except for lower tercile event forecasts which have value for the 30% decision threshold.

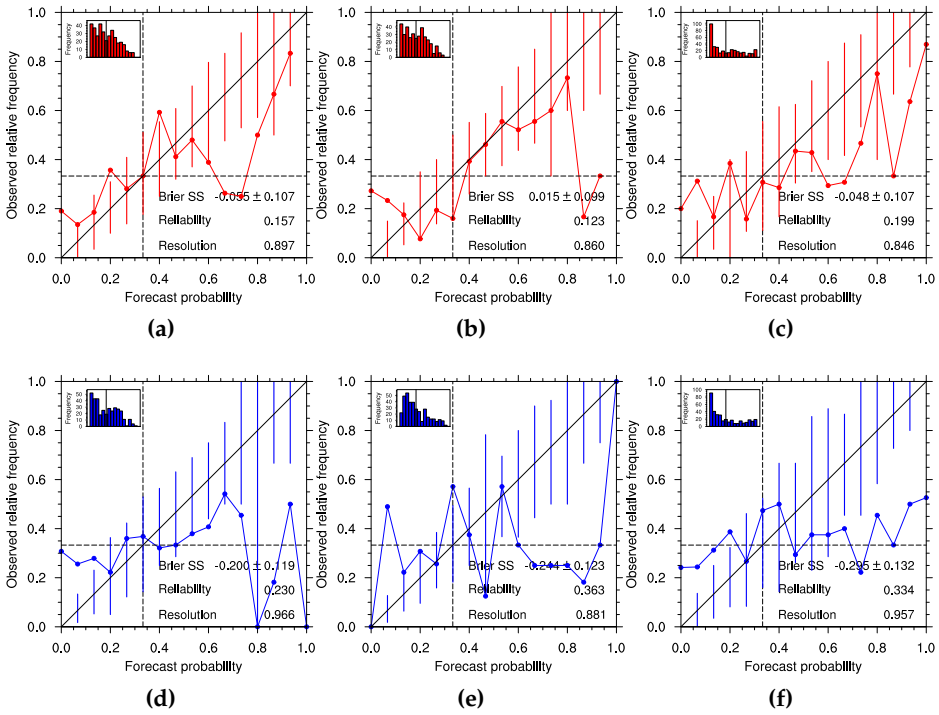


Figure C.21: Reliability of upper (a-c) and lower (d-f) tercile JJA temperature forecasts over Bangladesh. Reliability is shown for System 4 forecasts issued February (a & d), April (b & e) and June (c & f).

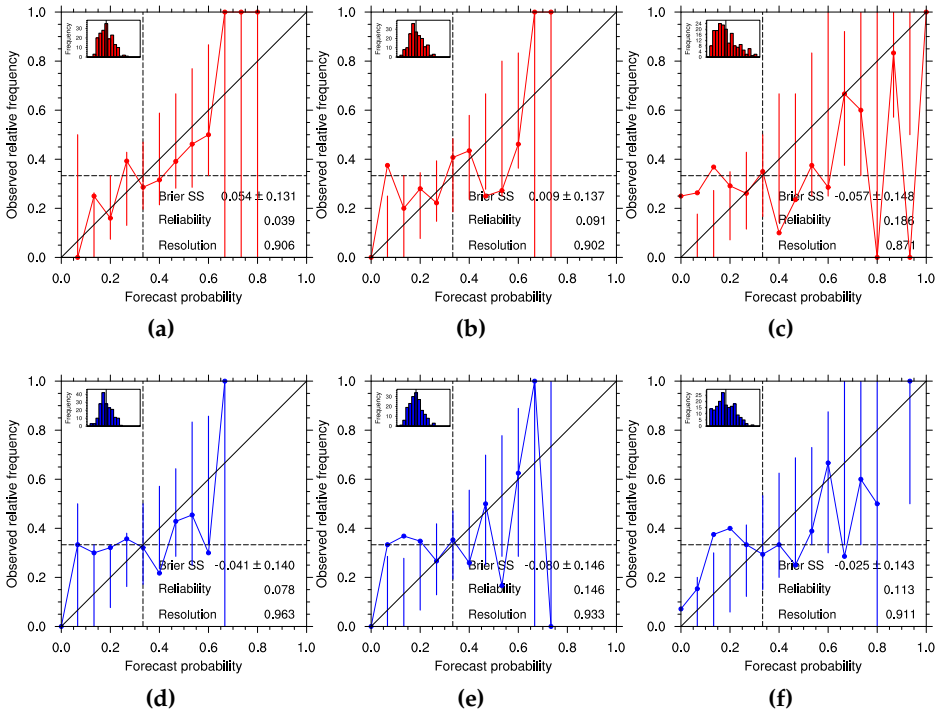


Figure C.22: Reliability of Bangladesh precipitation vs GPCP, details as in figure C.21.

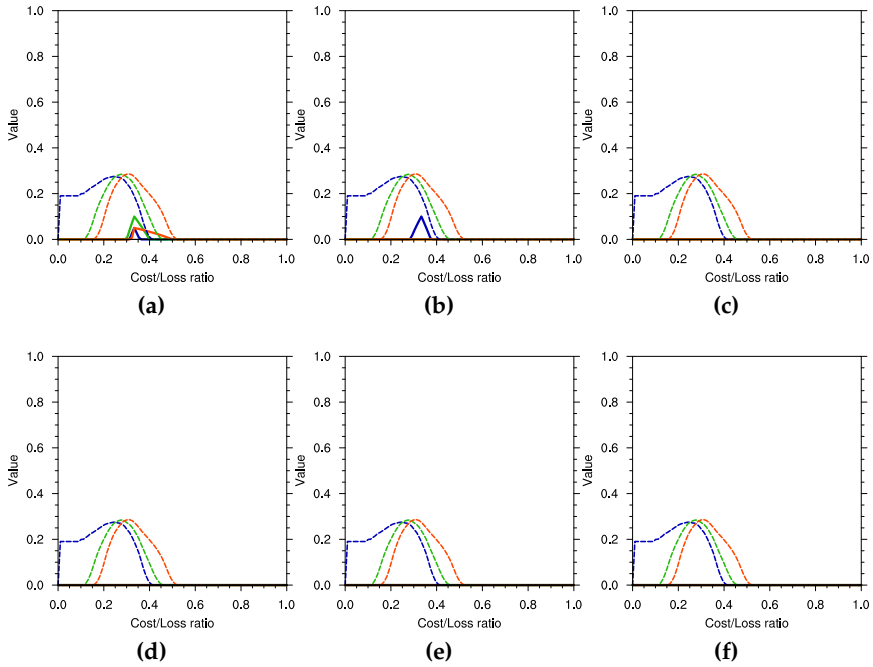


Figure C.23: Value of upper (a-c) and lower (d-f) tercile JAS temperature forecasts over Bangladesh, for System 4 forecasts issued February (a & d), April (b & e) and June (c & f). Curves for 30%, 50% and 70% decision thresholds are represented by blue, green and orange lines with dashed lines indicating 95% significance level for each threshold.

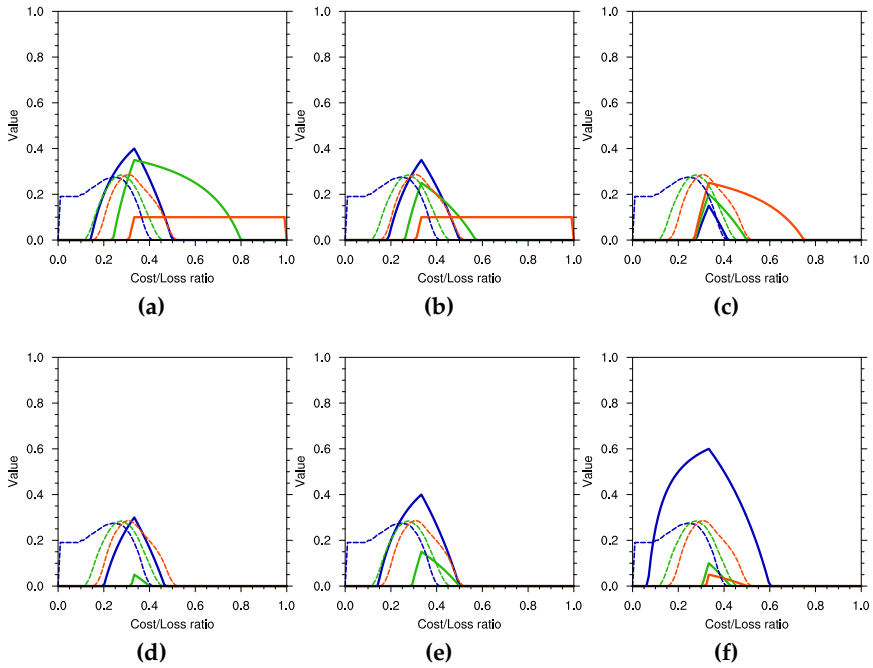


Figure C.24: Value of Bangladesh precipitation vs GPCP, details as in figure C.23.

APPENDIX D

Extra figures for Chapter 7

This appendix chapter contains extra figures corresponding to chapter 7, regarding the driving of the LMM with the ECMWF System 4 hindcasts. They have been separated from the main chapter for brevity. Results are discussed in the main chapter.

D.1 Appendix figures Botswana

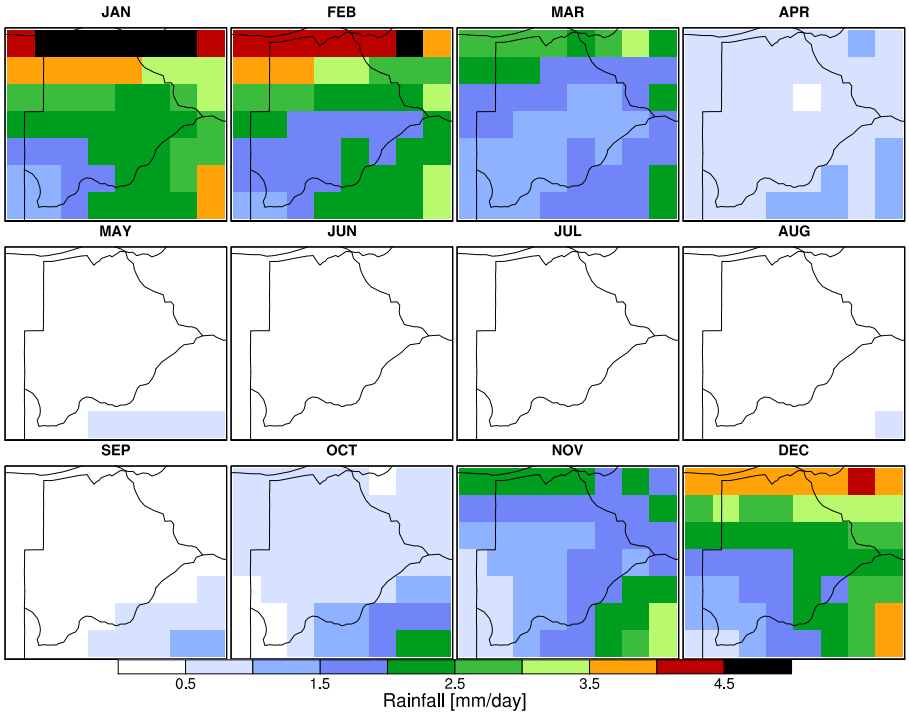


Figure D.1: ERA-Interim precipitation climatology map for Botswana.

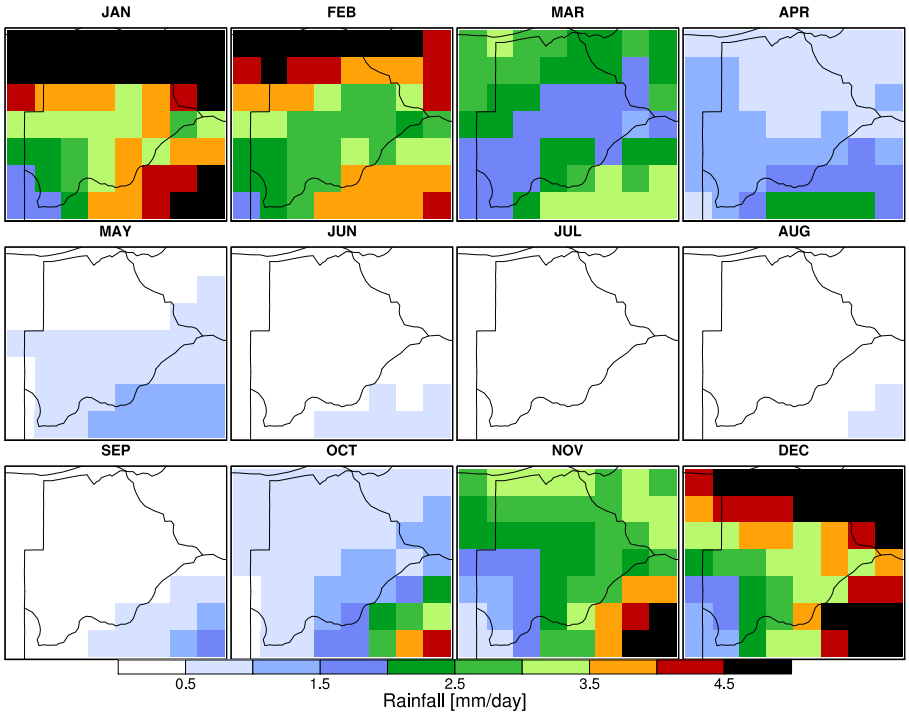


Figure D.2: System 4 precipitation climatology map for Botswana.

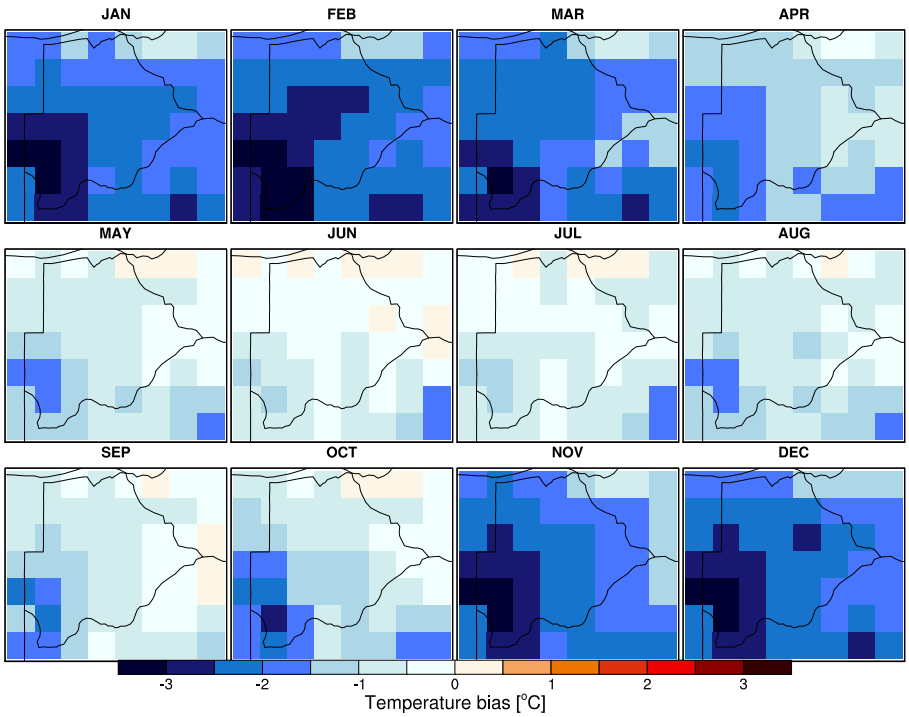


Figure D.3: System 4 temperature bias climatology map for Botswana (System 4 - ERA-Interim).

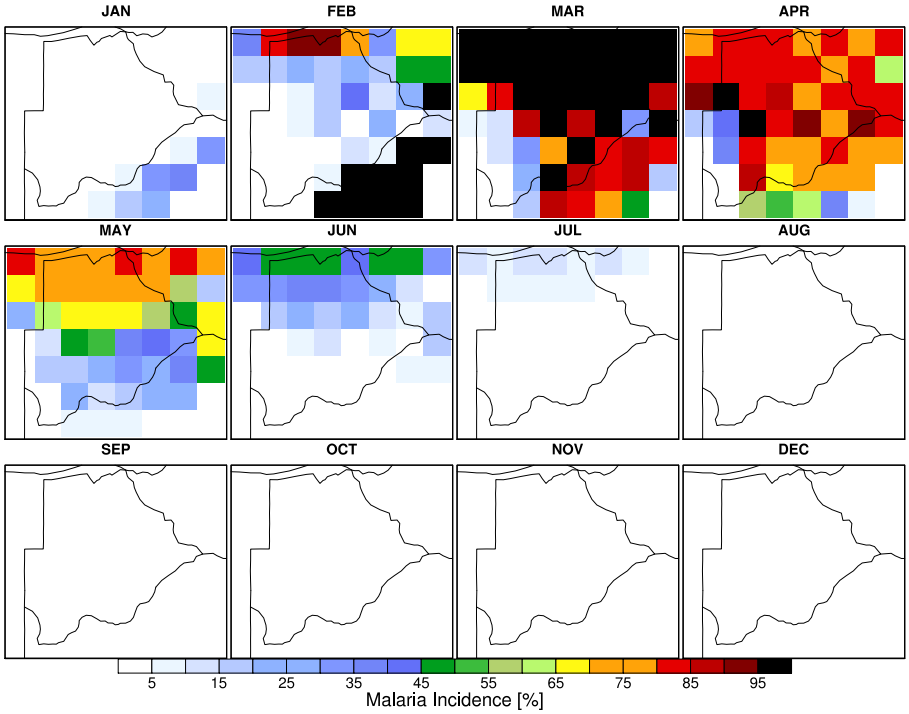


Figure D.4: ERA-Interim incidence climatology over Botswana.

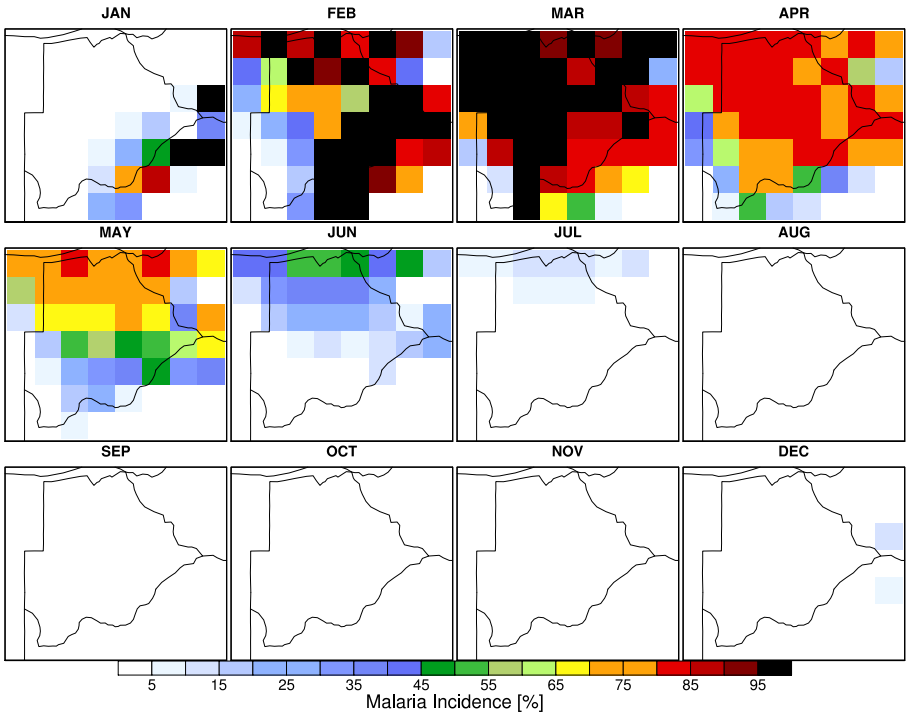


Figure D.5: System 4 incidence climatology over Botswana.

D.2 Sahel

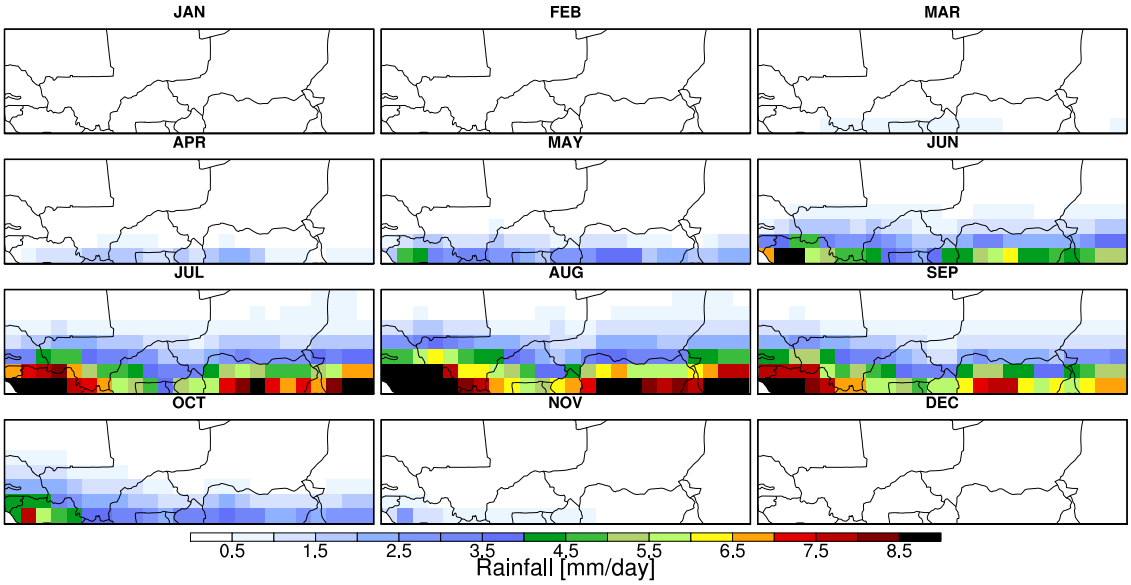


Figure D.6: ERA-Interim precipitation climatology map for the Sahel.

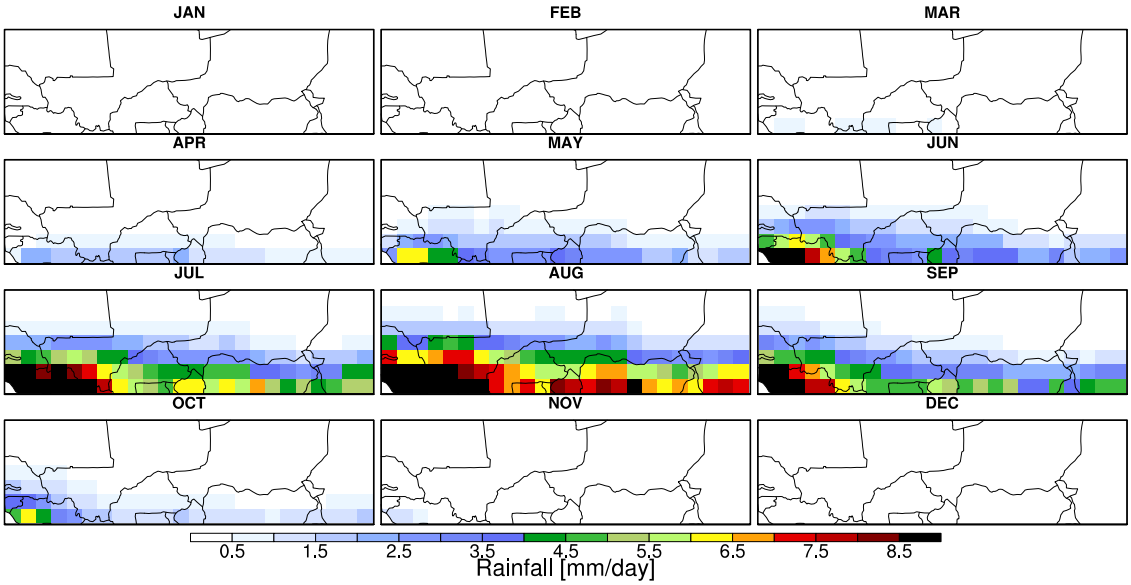


Figure D.7: System 4 precipitation climatology map for the Sahel.

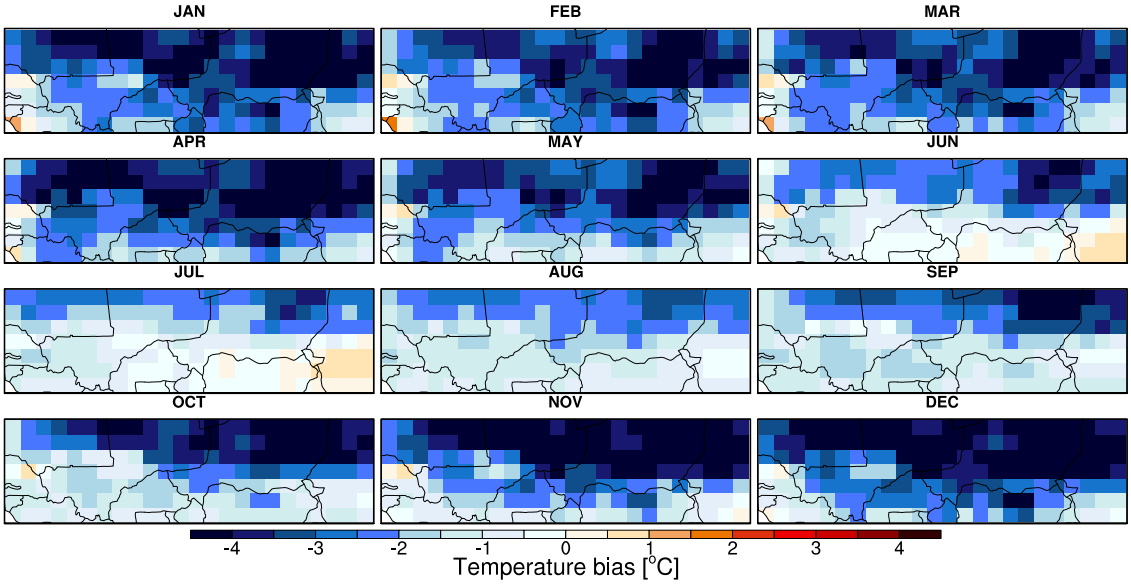


Figure D.8: System 4 temperature bias climatology map for the Sahel (System 4 minus ERA-Interim).

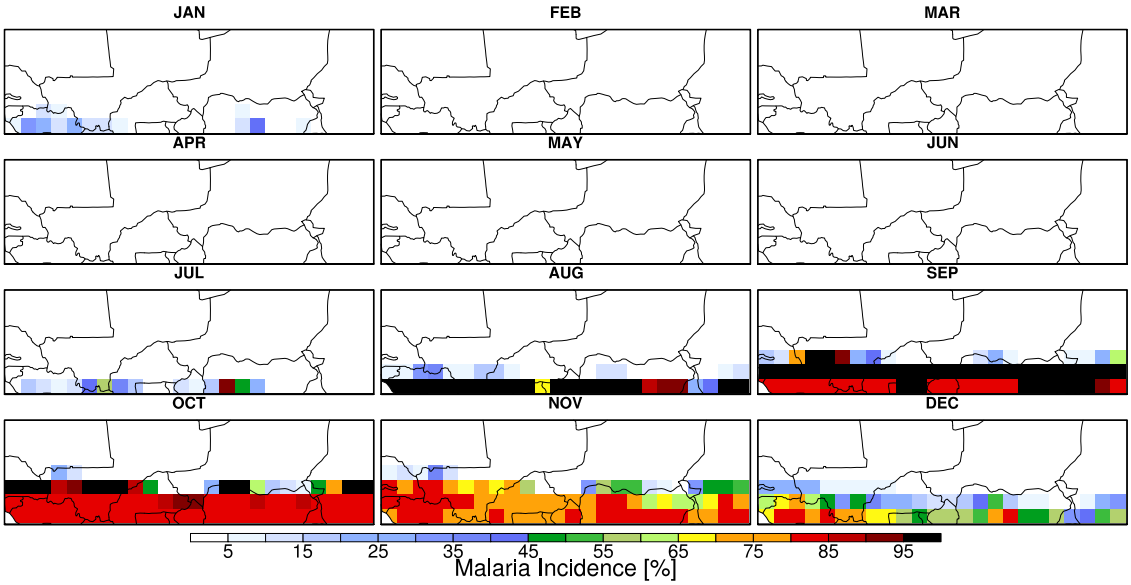


Figure D.9: ERA-Interim-driven LMM incidence climatology map for the Sahel.

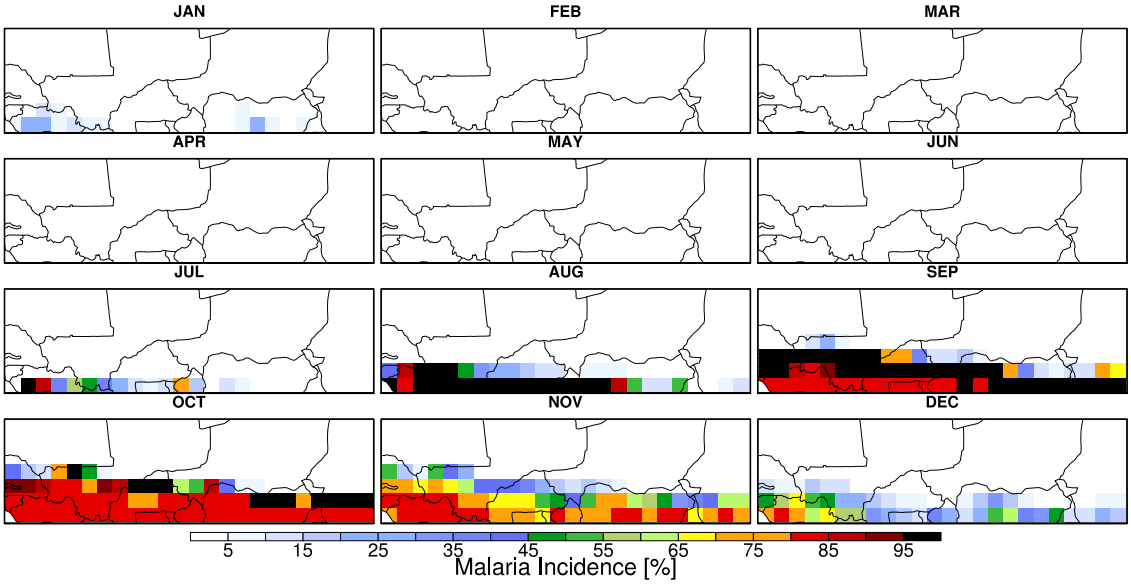


Figure D.10: System 4-driven LMM incidence climatology map for Sahel, with no bias correction.

D.3 Gulf of Guinea

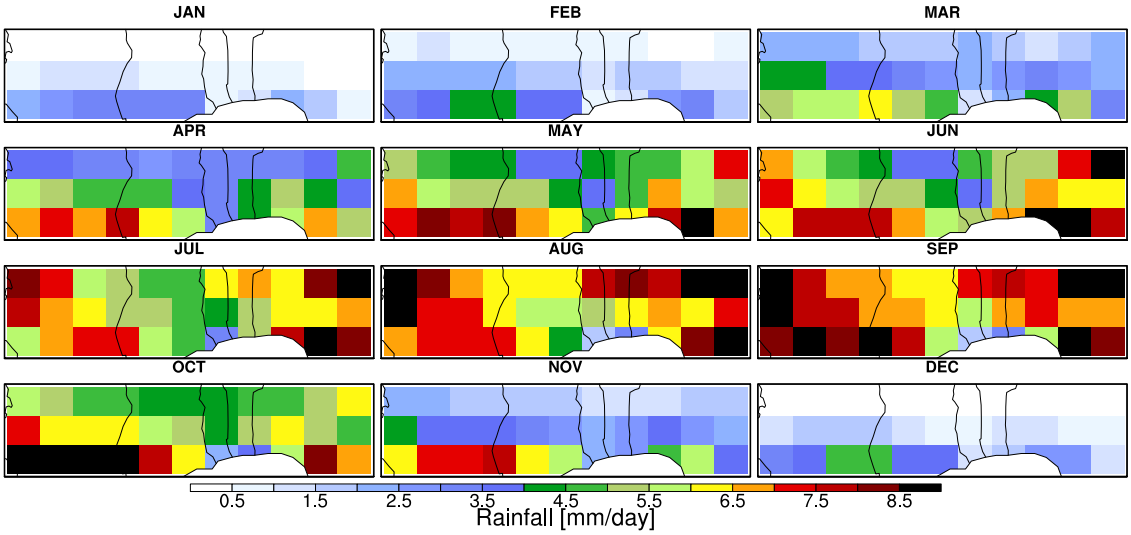


Figure D.11: ERA-Interim precipitation climatology over Gulf of Guinea.

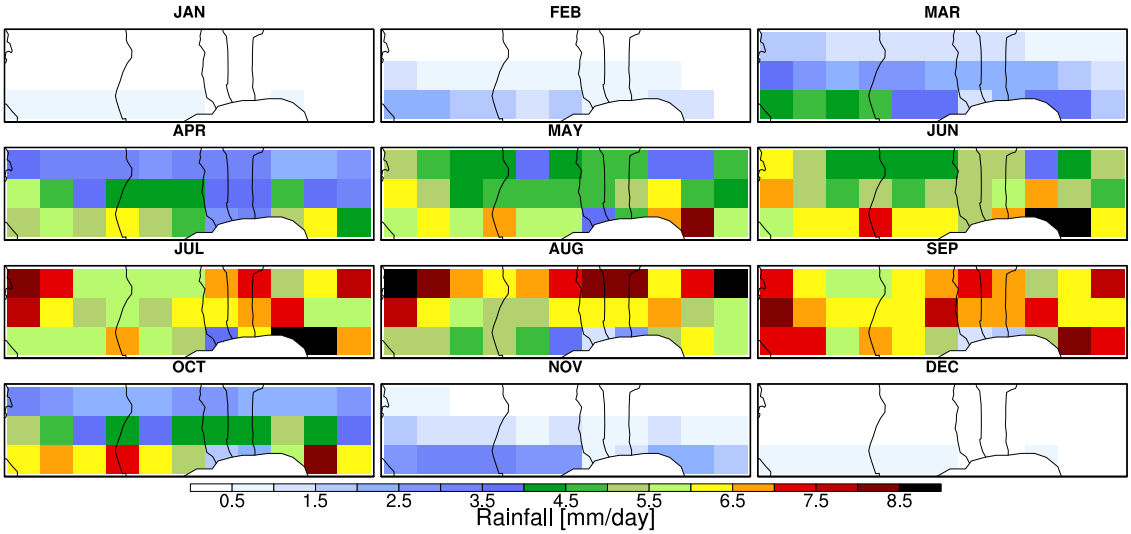


Figure D.12: System 4 precipitation climatology over Gulf of Guinea.

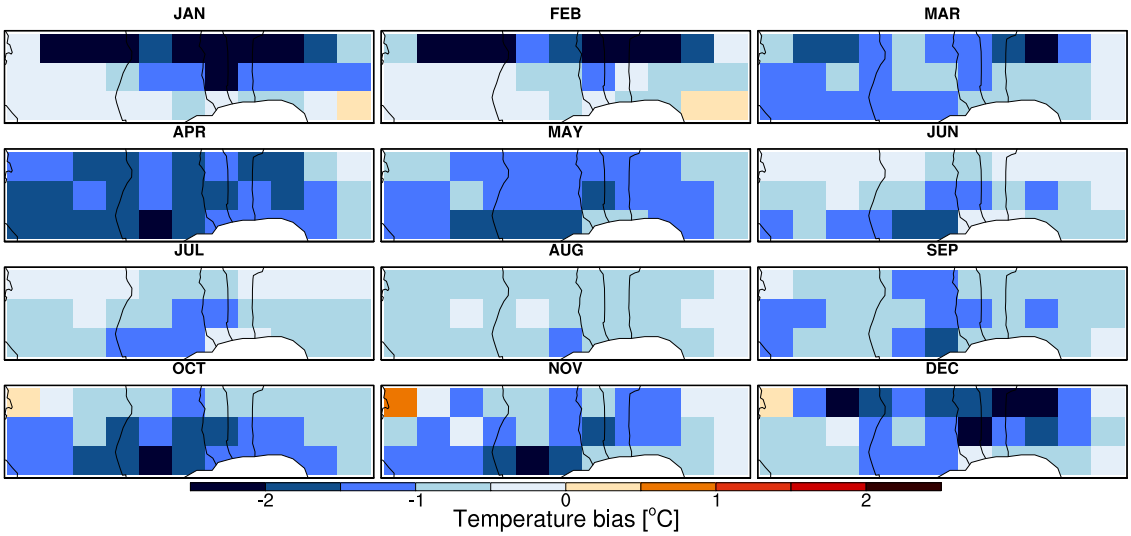


Figure D.13: System 4 temperature bias climatology over Gulf of Guinea (System 4 minus ERA-Interim).

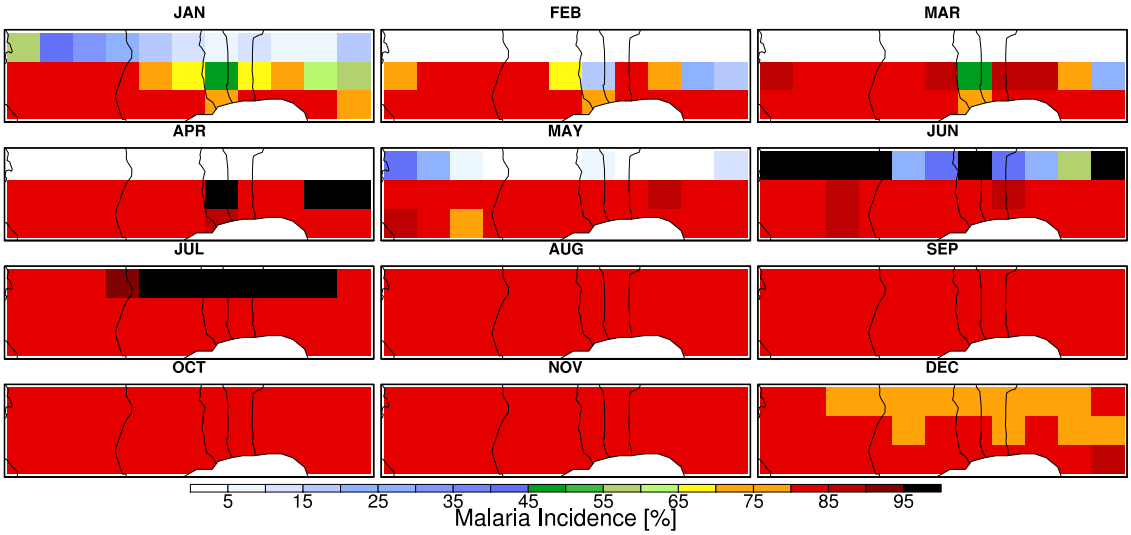


Figure D.14: ERA-Interim incidence climatology over the Gulf of Guinea.

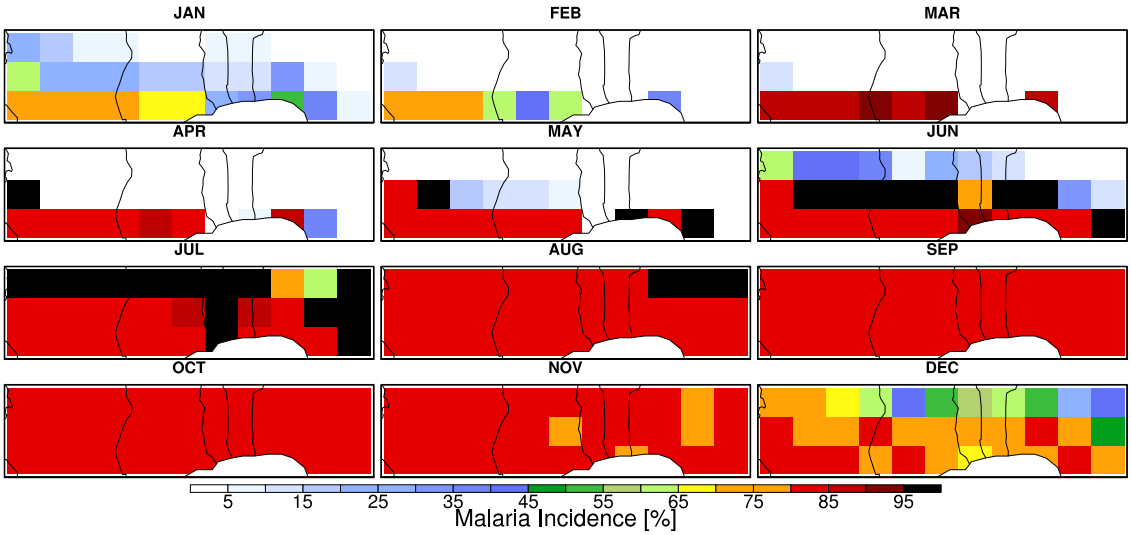


Figure D.15: System 4 incidence climatology over the Gulf of Guinea.

D.4 Malawi

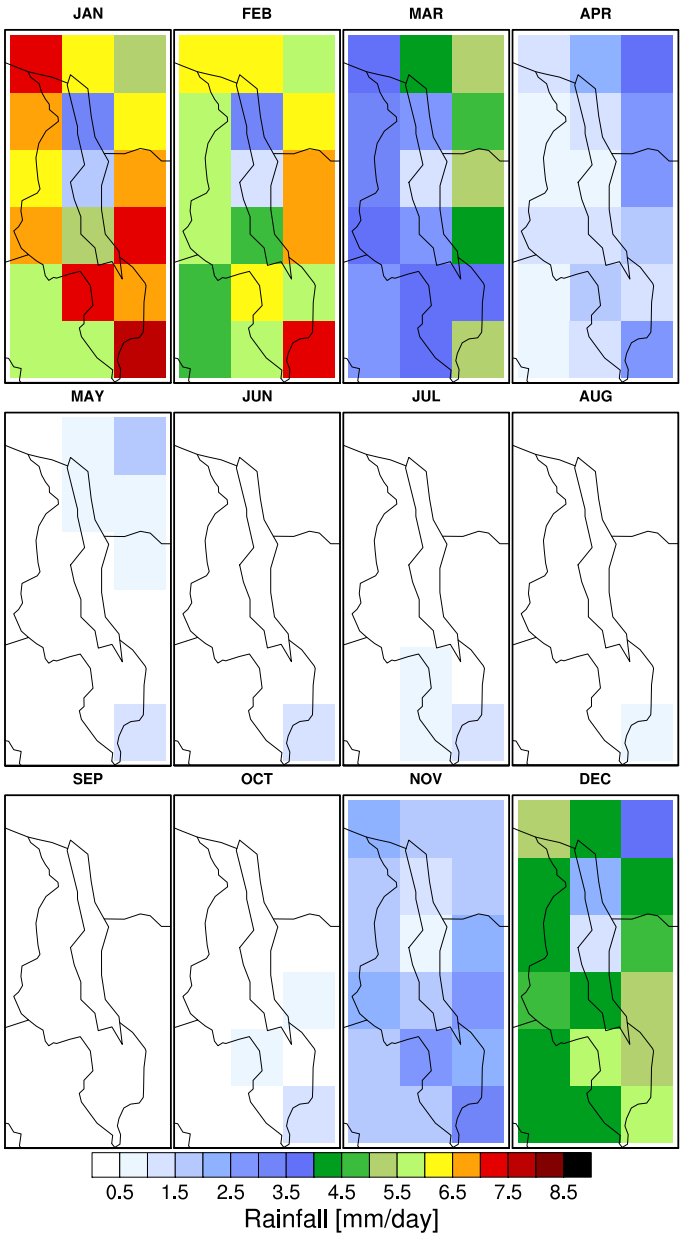


Figure D.16: ERA-Interim precipitation climatology map for Malawi.

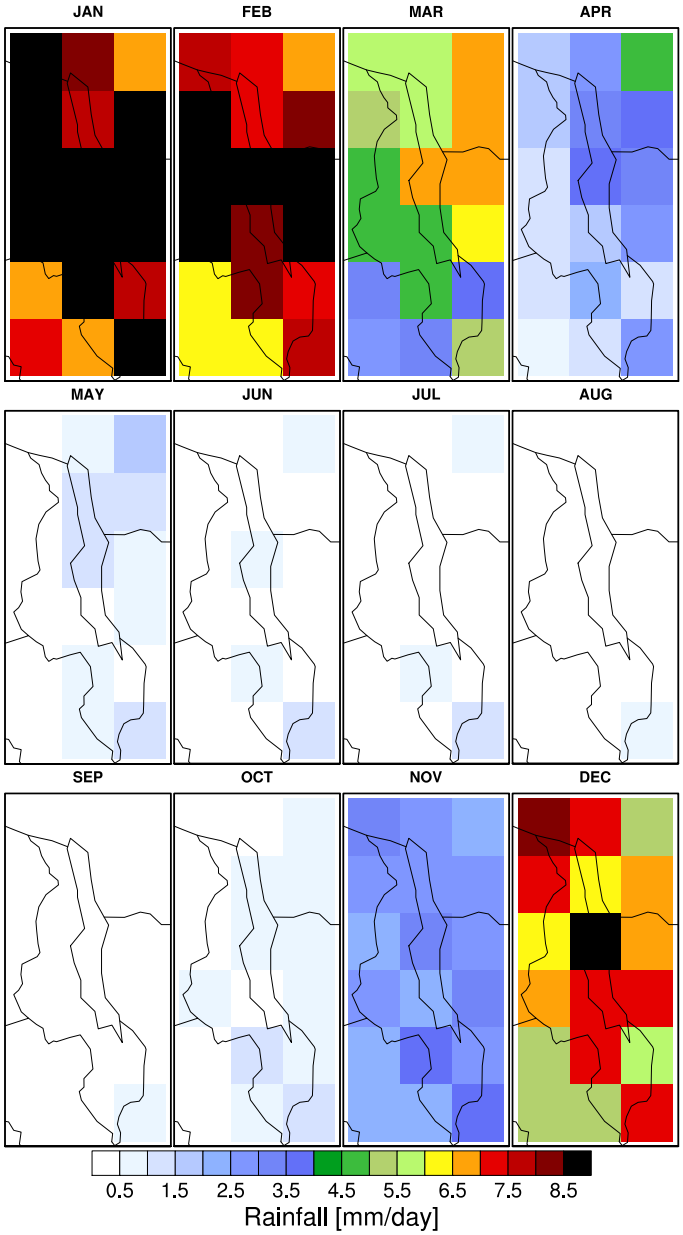


Figure D.17: System 4 precipitation climatology map for Malawi.

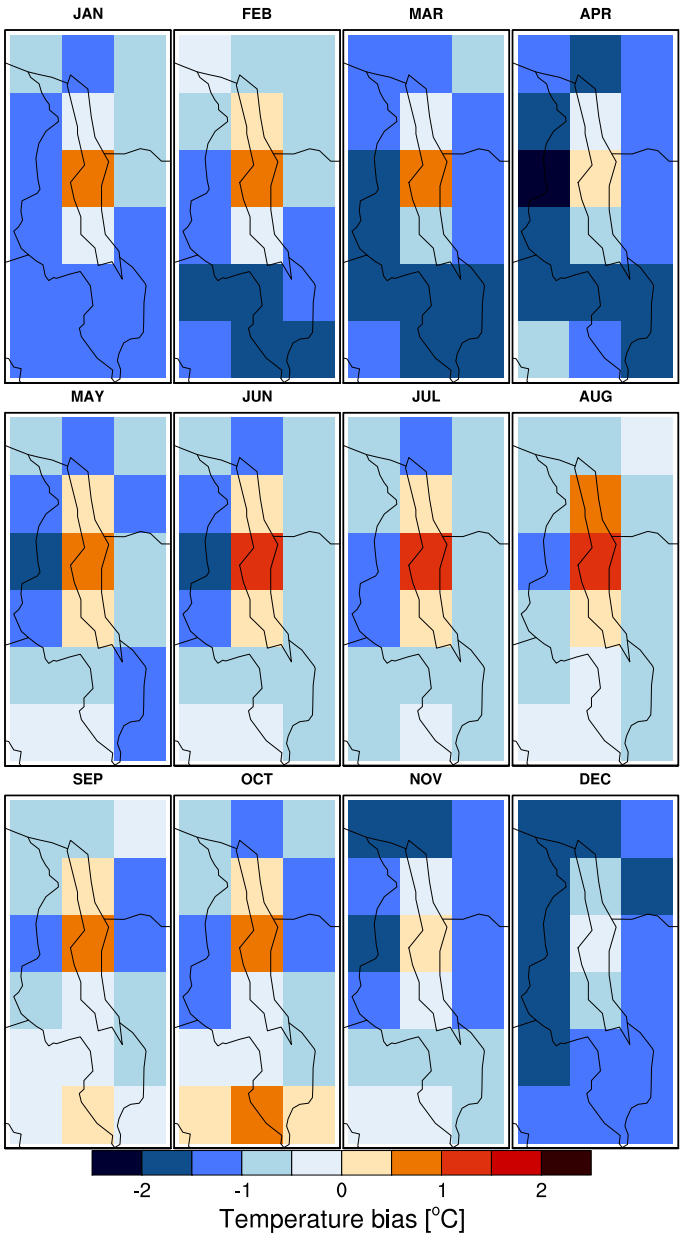


Figure D.18: System 4 temperature bias climatology map for Malawi (System 4 minus ERA-Interim).

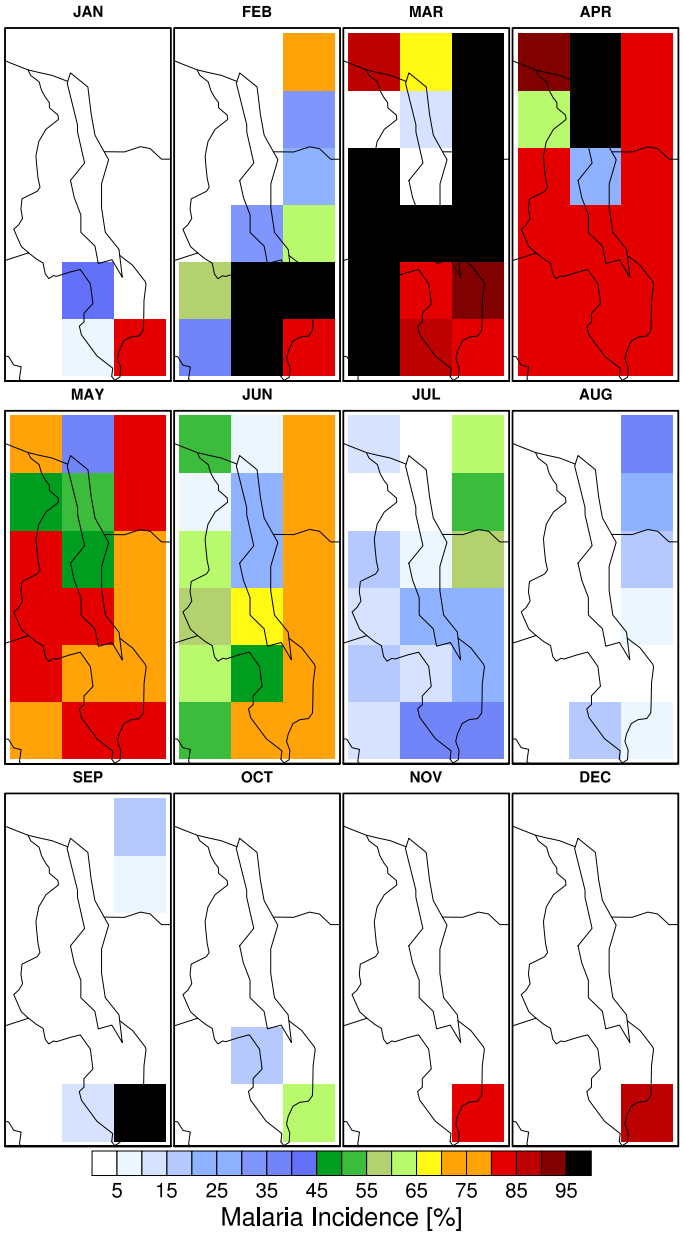


Figure D.19: ERA-Interim incidence climatology map over Malawi.

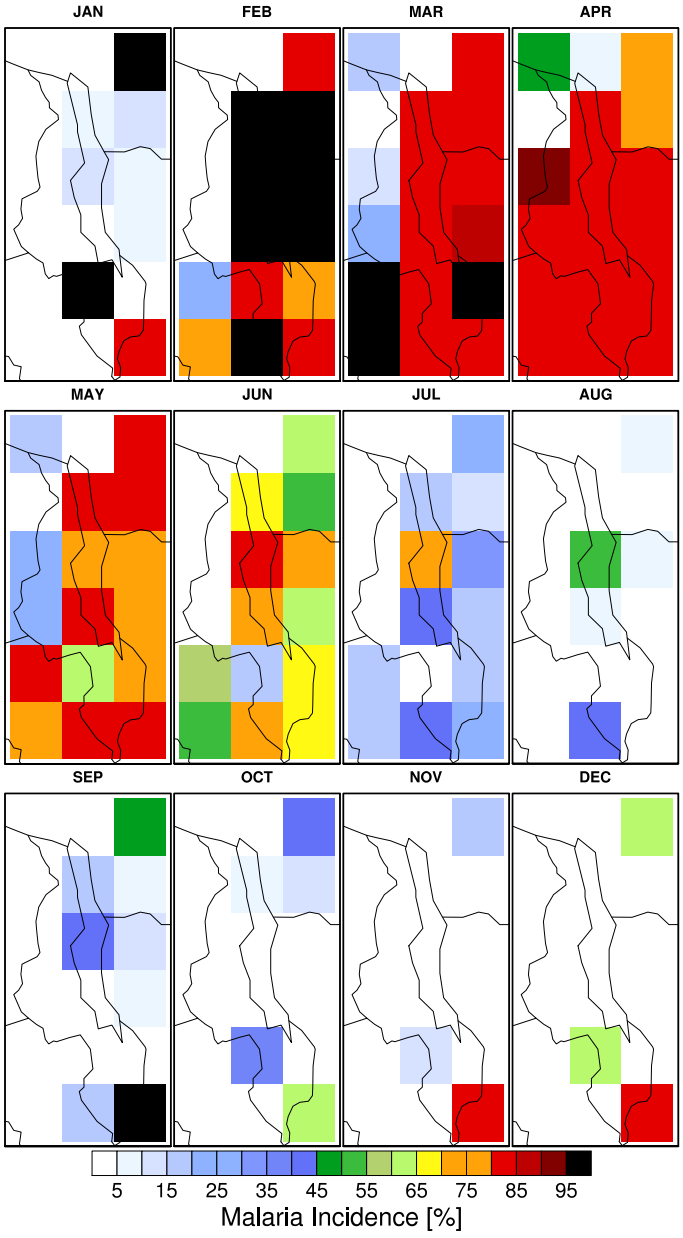


Figure D.20: System 4 incidence climatology map over Malawi.

APPENDIX E

Extra figures for Chapter 8

This appendix chapter contains extra figures corresponding to chapter 8. They have been separated from the main chapter for brevity. Only figures are contained in this appendix.

E.1 Scatter plots

E.1.1 Region B

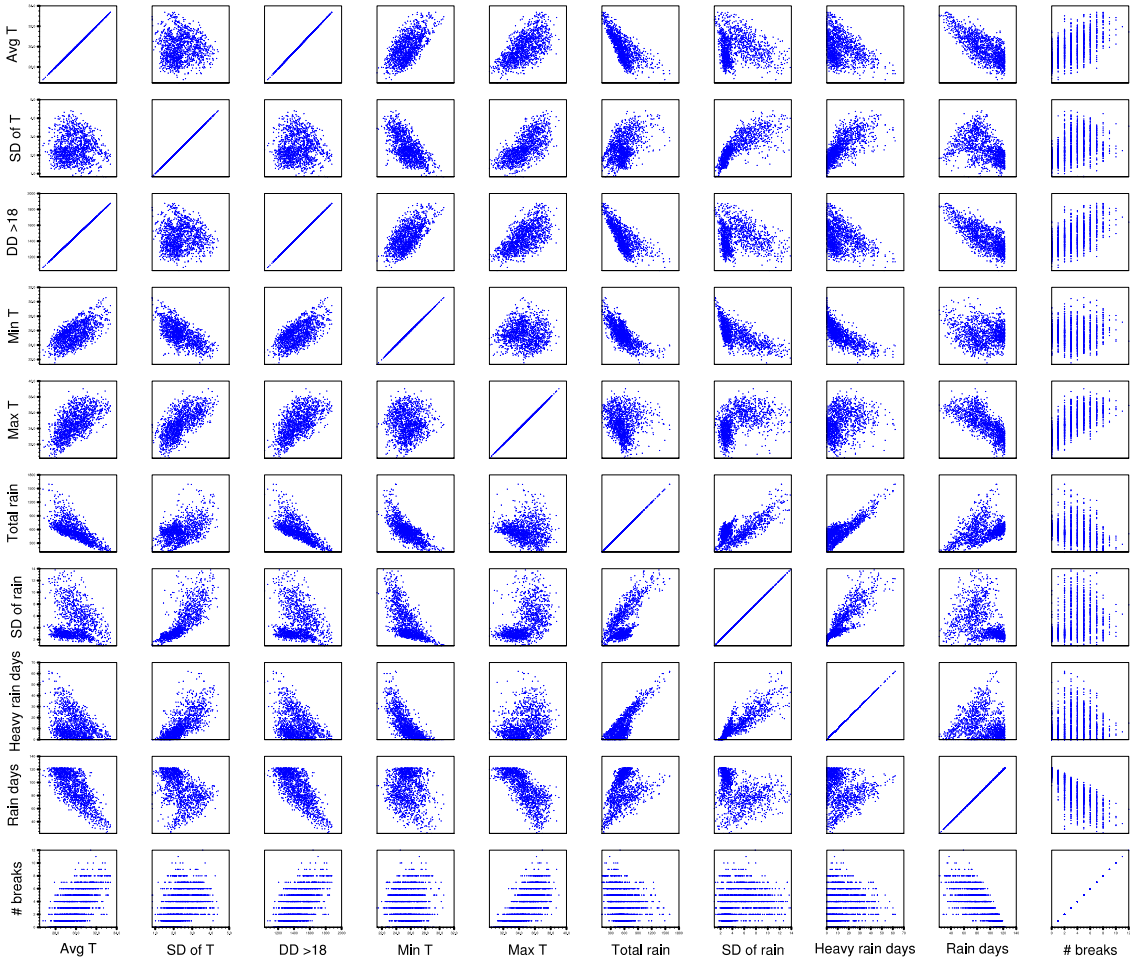


Figure E.1: Scatter plots for climate parameters vs climate parameters, JJAS, region B.

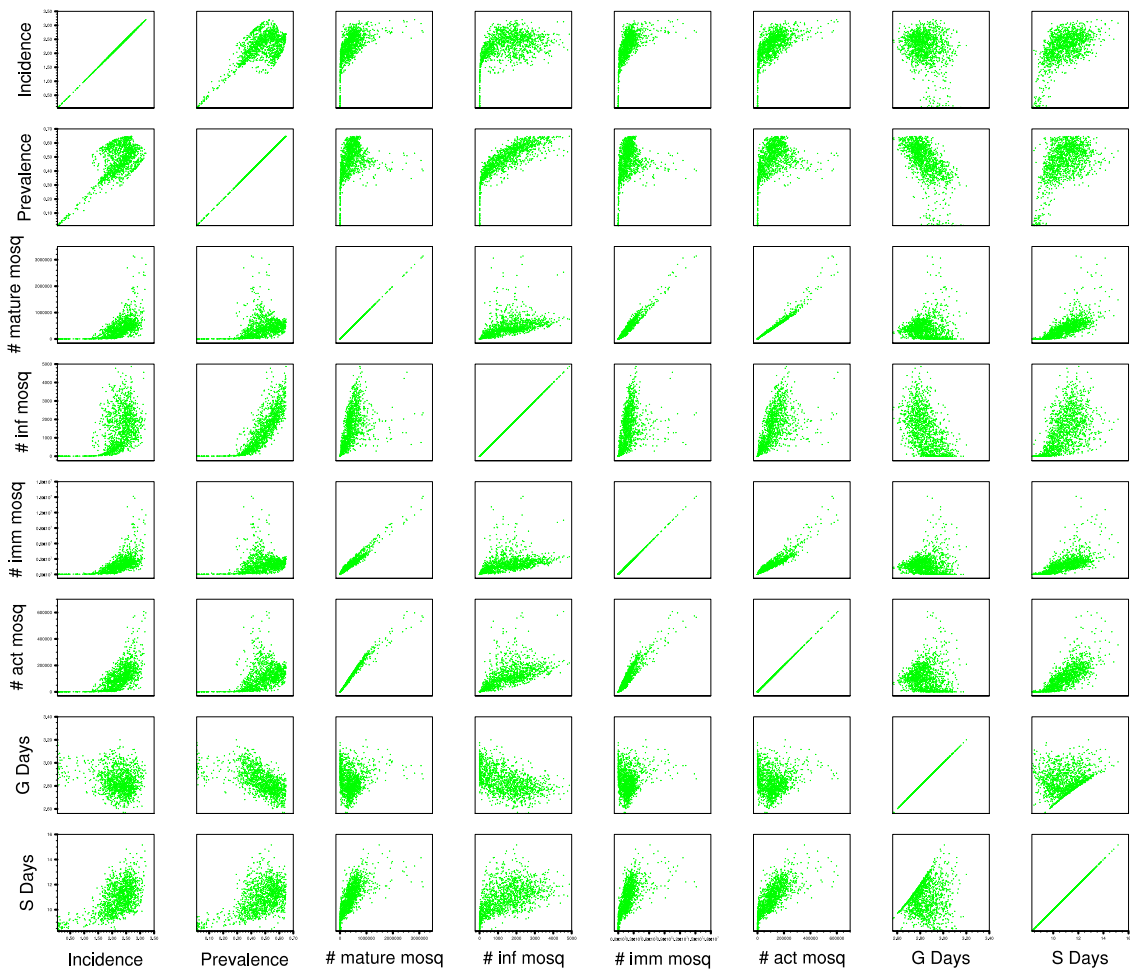


Figure E.2: Scatter plots for malaria parameters vs malaria parameters, JJAS/SOND, region B.

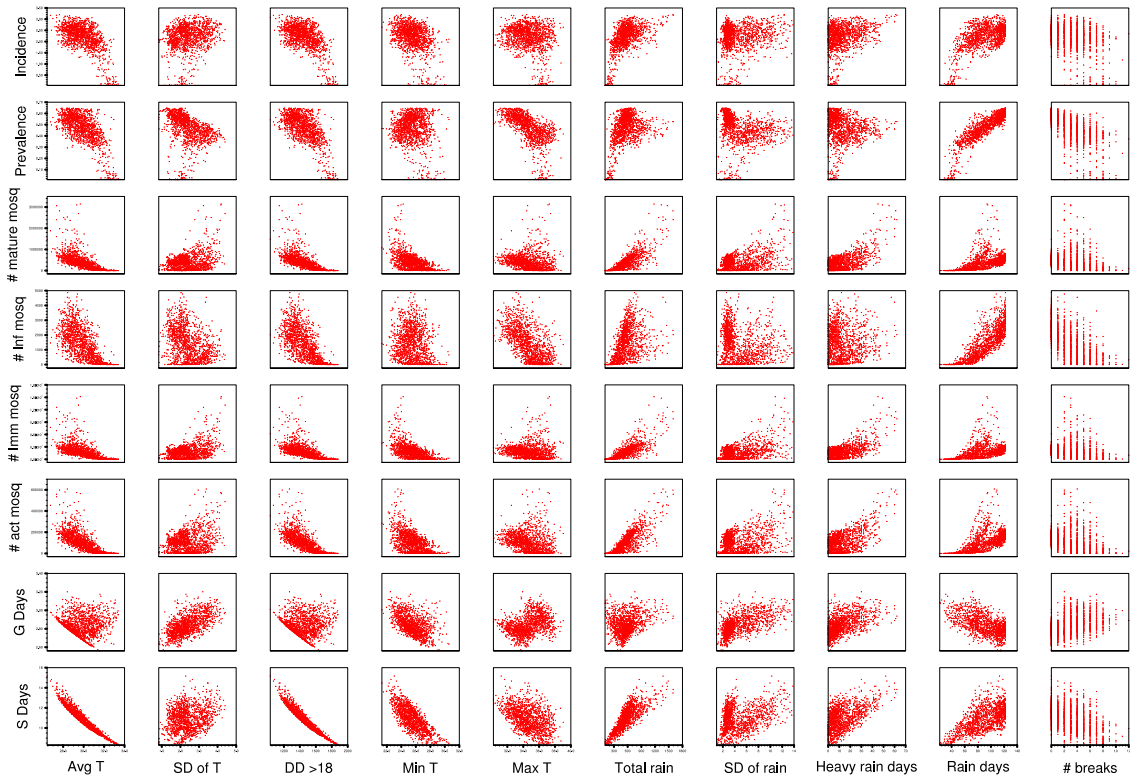


Figure E.3: Scatter plots for climate parameters vs malaria parameters, JJAS/SOND, region B.

E.1.2 Region C

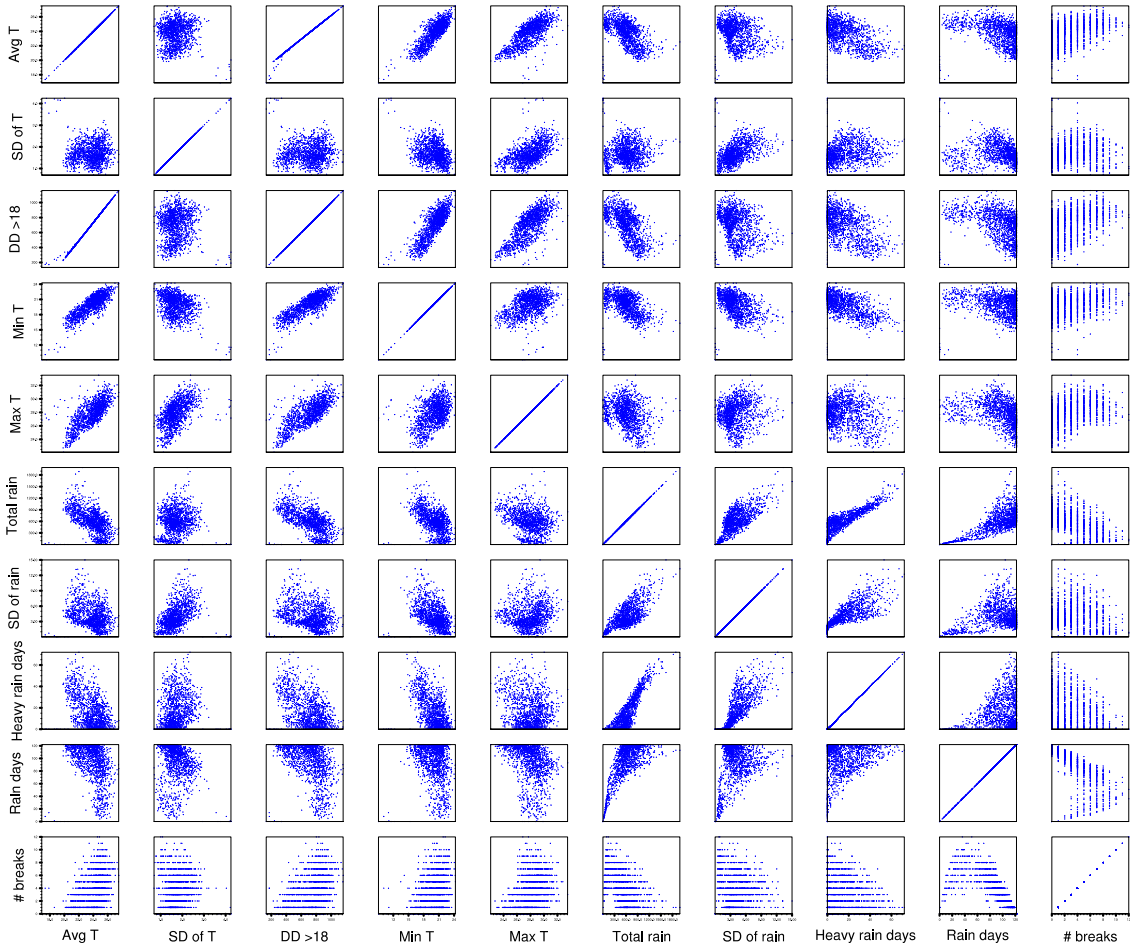


Figure E.4: Scatter plots for climate parameters vs climate parameters, JJAS, region C.

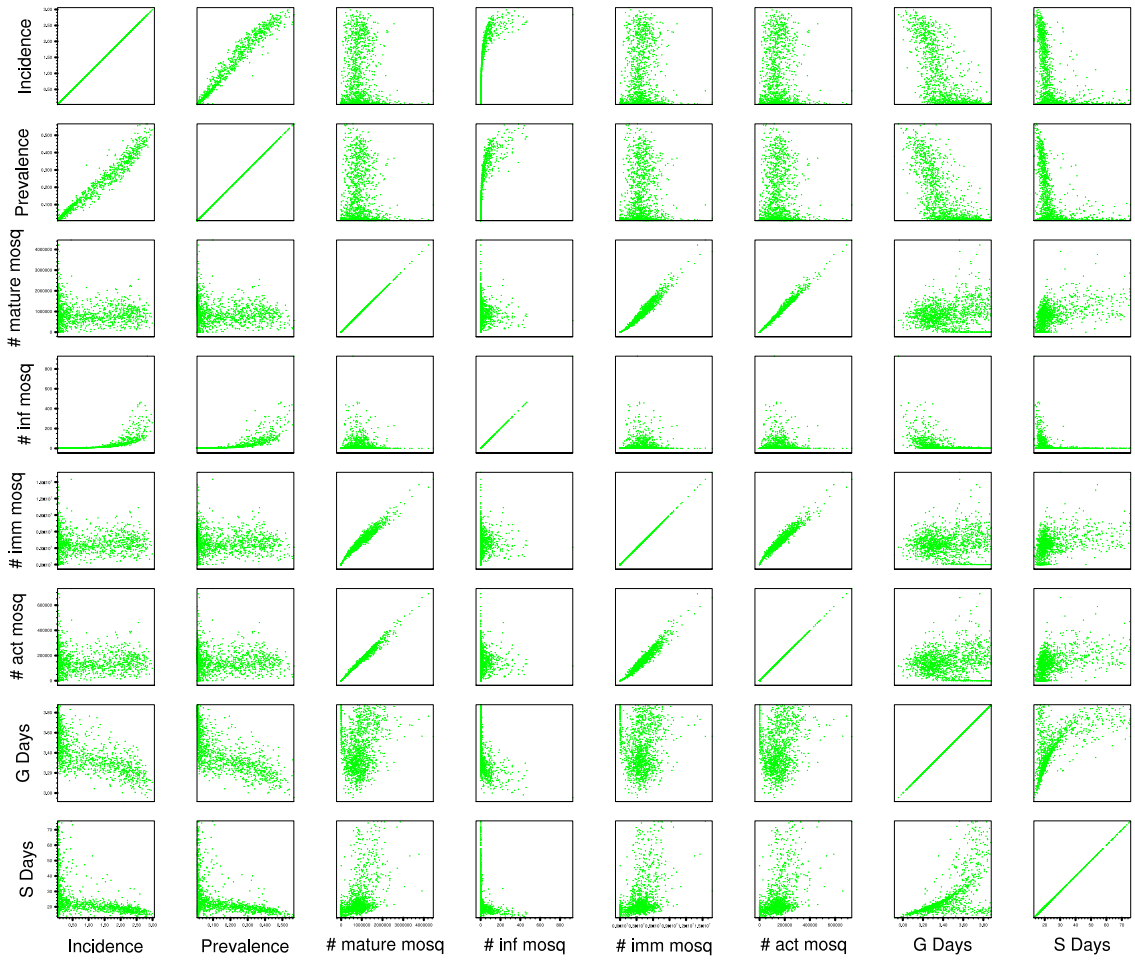


Figure E.5: Scatter plots for malaria parameters vs malaria parameters, JJAS/SOND, region C.

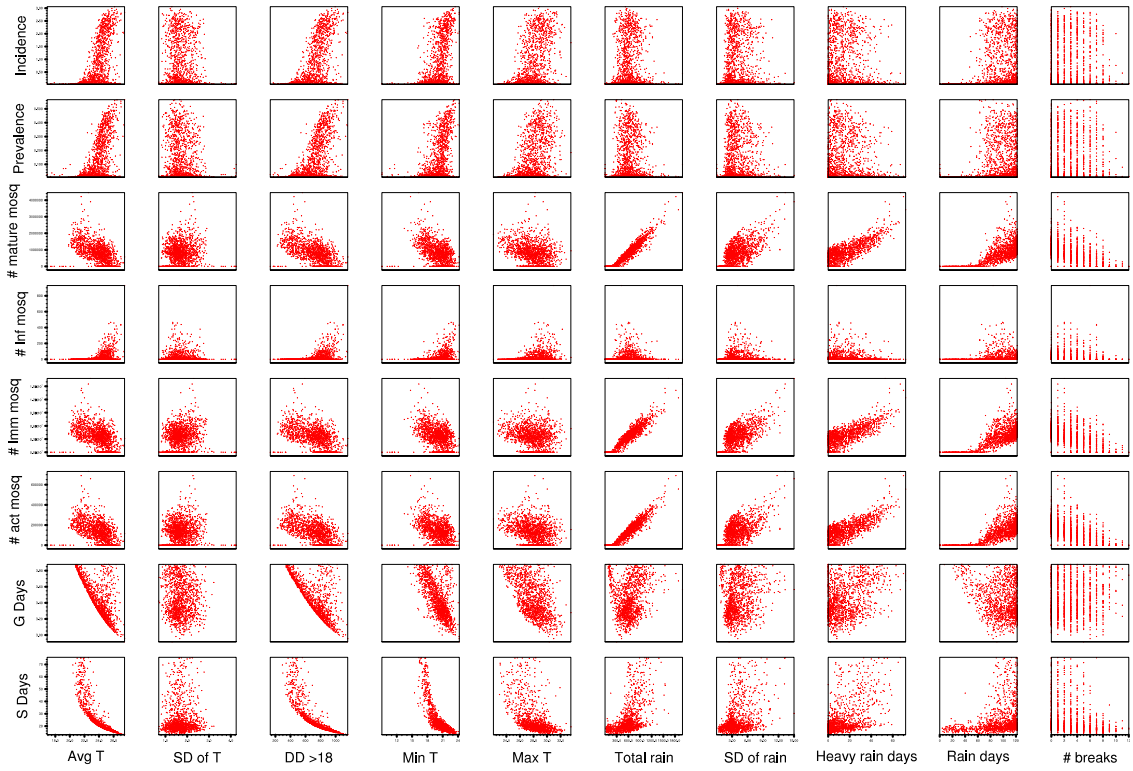


Figure E.6: Scatter plots for climate parameters vs malaria parameters, JJAS/SOND, region C.

E.2 Malaria seasonal cycle and impact surfaces, varying survival schemes.

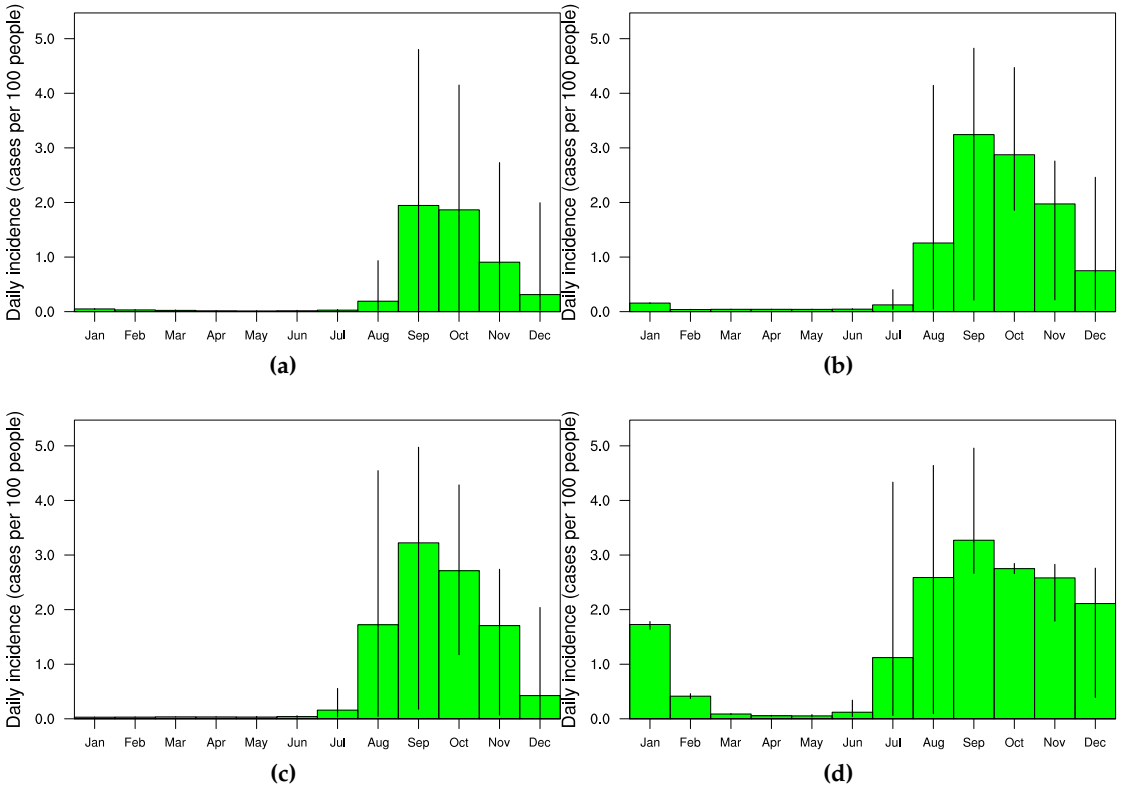


Figure E.7: Seasonal cycle of LMM incidence when driven by the 20th Century reanalysis, for region B, using survival schemes one to four (a-d).

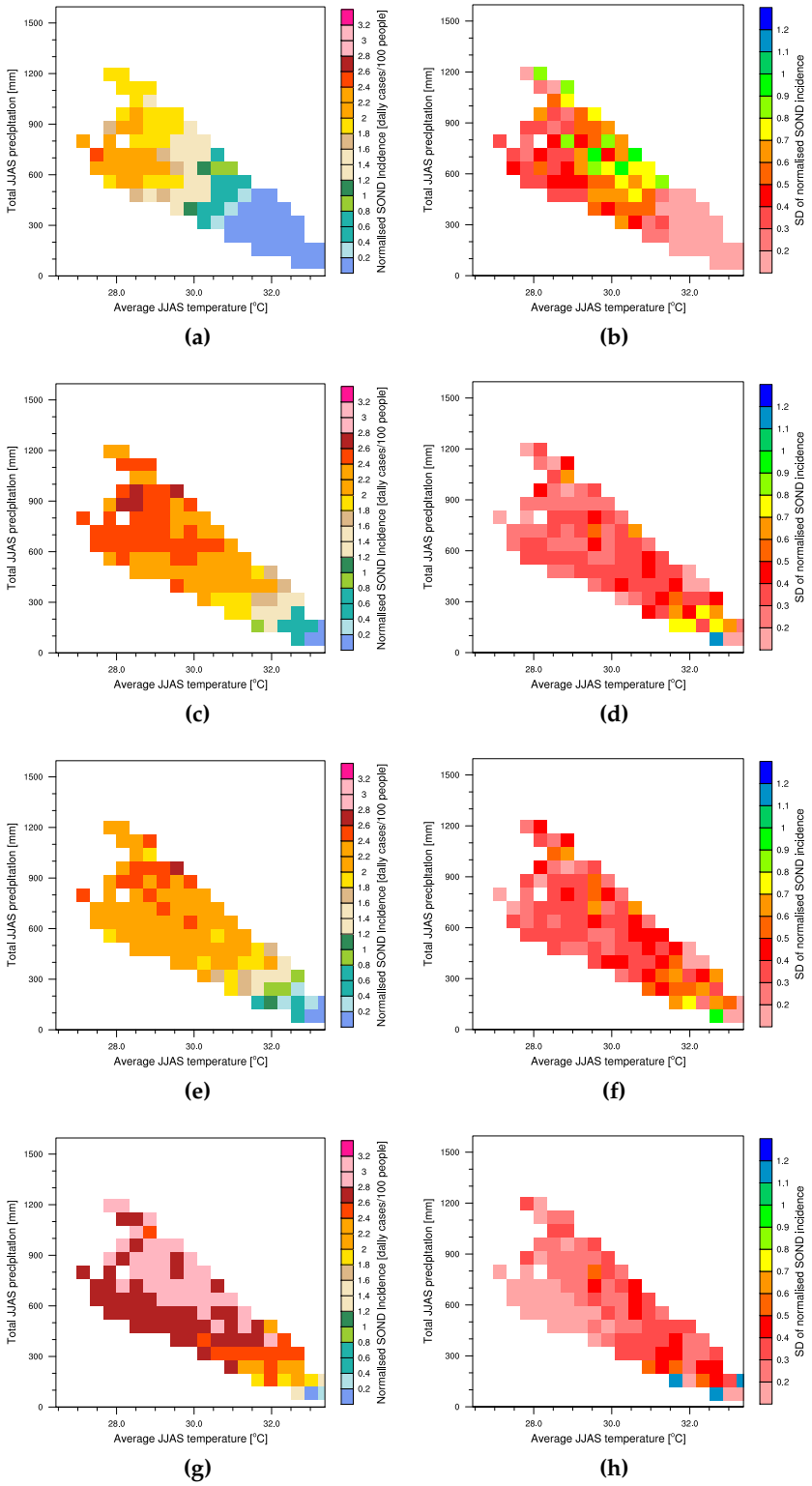


Figure E.8: Impact surface comparison, mean (left) and uncertainty (right) for region B. Using survival schemes one to four (top to bottom rows).

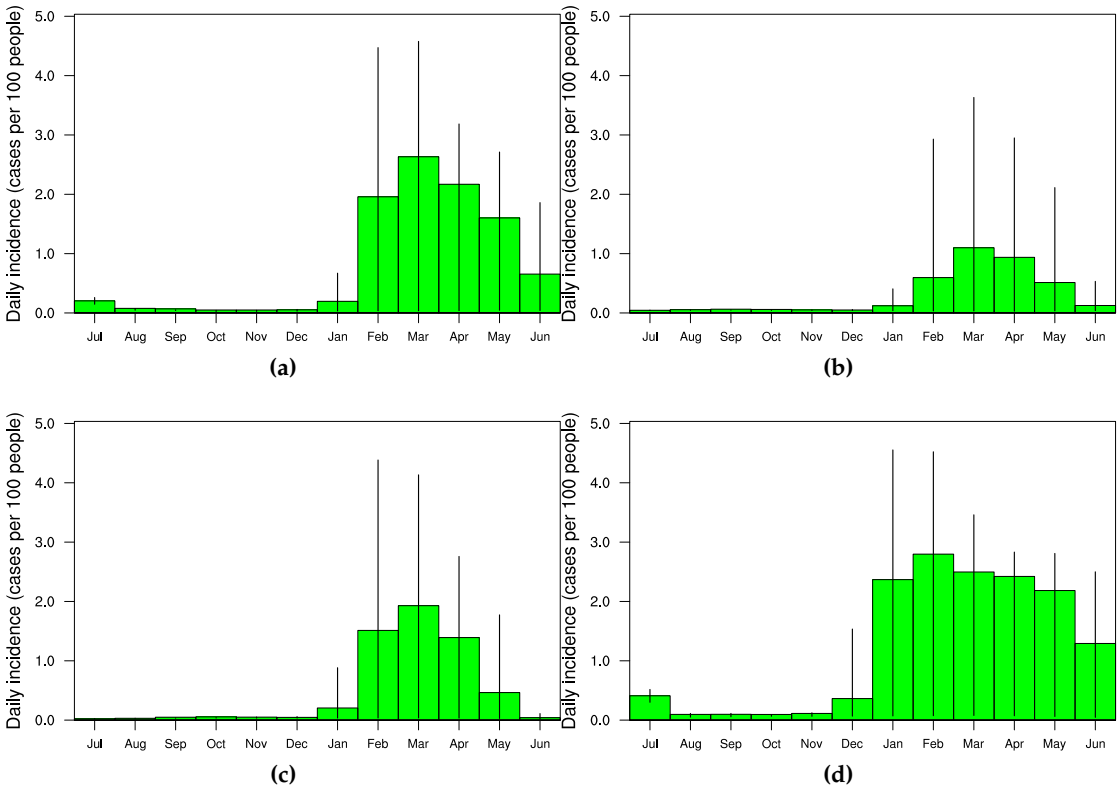


Figure E.9: Seasonal cycle of LMM incidence when driven by the 20th Century reanalysis, for region C, using survival schemes one to four (a-d).

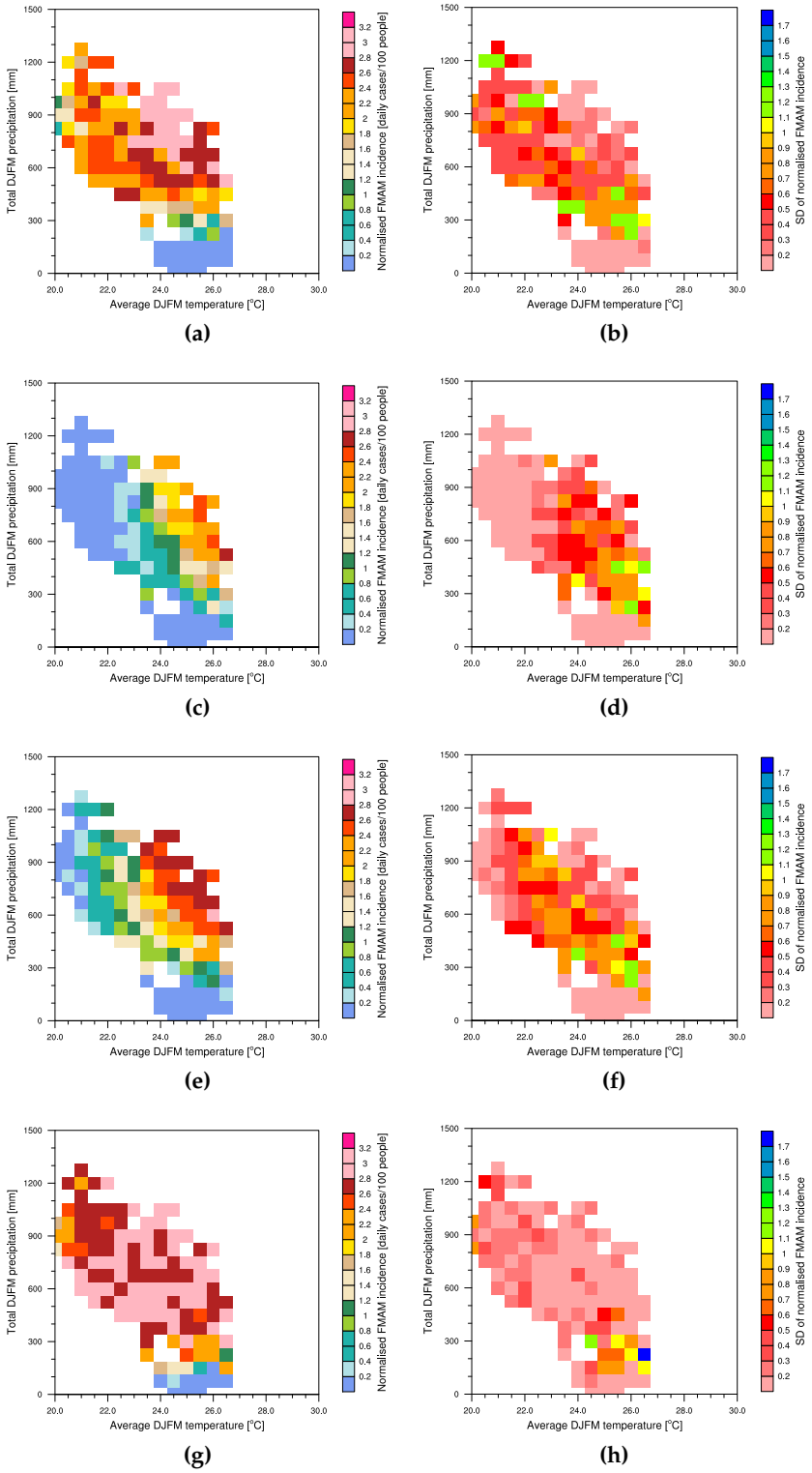


Figure E.10: Impact surface comparison, mean (left) and uncertainty (right) for region C. Using survival schemes one to four (top to bottom rows).